Influencing factors for toxicity of water reservoirs

Akshay Deshpande
Computer Science Dept.
Clemson University
Clemson, USA
deshpa2@clemson.edu

Chaitanya Mundle
Computer Science Dept.
Clemson University
Clemson, USA
cmundle@clemson.edu

Poojitha Tarasani
Computer Science Dept.
Clemson University
Clemson, USA
ptarasa@clemson.edu

Srivathsan Mohan
Computer Science Dept.
Clemson University
Clemson, USA
srivath@clemson.edu

Supreeth Ramakrishna
Computer Science Dept.
Clemson University
Clemson, USA
supreer@clemson.edu

***Abstract*—Factors influencing water quality have been significant environmental issues in the current generation. Several components contribute to the toxicity of groundwater resources such as lakes, ponds, rivers, and other water bodies. Due to the exploitation of resources in the name of industrialization, large corporations have polluted various water bodies without proper regulation. Many communities residing in these regions are forced to consume water of degraded quality because of a lack of alternatives. In this paper, we aim to figure out the reasons for the deterioration of water quality by crunching massive datasets that contain data about water composition and demographics information for different regions in South Carolina. We also analyzed such various types of files which gave information about various industries present in that geographic location. Using different scalable machine learning algorithms for these vast source files, we try to prove the correlation among these factors that will show the bias that may exist among other neighborhoods concerning water contamination.**

***Index Terms*—Specific Conductance,South Carolina Water Reservoirs, United States Geological Survey, ,Toxicity, Toxicity Release Inventory Data, Undersampling, Oversampling, SMOTE, Dissolved Oxygen, Decision Tree Classifier**

I. Introduction

Water quality control is one of the most relevant issues our communities are facing in the current world. Research team in USGS (United States Geological Survey) probes water bodies for toxicity across the United States of America. However, in this project we would be considering data from the state of South Carolina as a sample for our analysis. Due to huge industries, large water reservoirs for humans (and animals) are becoming more toxic and are not of drinking quality . People are hesitant to drink water from these resources, but can not afford any other alternative.We want to figure out what makes the water quality[3] bad and ascertain if there is any correlation between the toxicity of water and the factors that influence the toxicity along with the type of neighborhood and the demographics in the neighborhood. In order to verify if there is any correlation between the above mentioned factors, we need to build explainable algorithms that show bias that may exist with respect to water contamination among different neighborhoods.[9]

M. Tiemann. Safe drinking water act (SDWA): A summary of the act and its major requirements. Congressional Research Service Washington, DC, 2014.We would be performing analysis on the Water composition data we have obtained from the USGS (United States Geological Survey) research. Using this dataset, we would be leveraging measurements such as PH-value, Dissolved Oxygen levels, Specific Conductance etc.[10] in order to get the details about the extent of contamination in individual water reservoirs across South Carolina from 1990 to 2020. We would also be using datasets that we have obtained from Census data obtained from the U.S. Census Bureau, along with Healthcare dataset obtained from the research conducted by Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute. The initial process would be majorly focused on determining the contamination which is based on the Dissolved oxygen data. This is purely based on the lead which acts as a catalyst in water contamination. This fact was taken into consideration by authors Jacob Abernethy, Jared Webb, et al. wherein they tried to predict the lead contamination by each house-hold's water supply within a city[8]. A similar study of determining the water quality was conducted by Jachimowski, A. In the analysis the supply zones of four municipal water treatment plants in Krakow were considered. The selection of 29 water sampling points within the supply area allowed comparing water quality with respect to operational and technological aspects. This analysis of factor sanctioned four major components which were responsbile for correlations between the tap water quality variables for distinguishing which fleshed out a total of 77% variability of water.[11]

Applying advanced machine learning algorithms on the combination of these huge datasets, we would get results that intelligently identify patterns such as any kind of correlation between the contamination levels in the water reservoirs and other and the factors that influence the toxicity. Our model can identify if there exists any skewed contamination levels across

various types of neighborhoods.

• In this paper, we address the problem of identification of toxics in the water reservoir in the state of South Carolina and classify water into two classes as pure or contaminated water using Machine Learning Algorithms.

• In order to classify the class to which the water belongs to we take into consideration two major factors i.e Specific Conductance and Dissolved Oxygen

• This classification is further complemented by analyzing patterns in the Chemical releases by Industries in South Carolina to determine the Facilities releasing various toxics into water

• The Pattern Analysis helps to determine Lead and Lead Compounds as the major chemicals released by Facilities into the water reservoirs of South Carolina leading to Water Contamination over the years.[7]

• The Correlation between the Specific Conductance and Dissolved Oxygen in water and the release of Lead and Lead Compounds is determined and the top 7 Facilities contributing to water Contamination are identified[13]

II. APPROACH

A. Data Scraping

This type of data is not specifically used for project purposes. However, analyzing such a dataset would be a bit of a task and hence would serve the purpose as a whole of working on a complex and unstructured data. Hence, according to these attributes we specifically opted for the USGS(United States Geological Survey) water data which was scraped via the use of internet from the official site of USGS water department. This dataset fleshes out some of the most important factors which are responsible for the water contamination across South Carolina over the years. The scraping was done via searching and sorting multiple datasets with different domains. A constraint here was to convert and scale the data into a more readable format. This is because the original format was ".out" and after we ran the scripting commands in python, the data was in a comma separated values (csv) format. Now, as the size of the data was huge, to read it using pandas was a task. This is why we decomposed the original data into multiple subsets and implemented the analysis on each subset.[2]

1) Main Dataset:

- Source: United States Geological Survey
- Link : https://waterdata.usgs.gov/sc/nwis/qw
- Size of the Dataset: 5.5 GB
- Number of features: 11 columns
- Number of records: More than 100 Million rows
- TimeLine: 1990 to 2020
- Key Features: Temperature, Specific Conductance, Dissolved Oxygen, PH Value, Precipitation, Gage, height,Timestamp

2) Supporting Datasets: The Toxics Release Inventory(TRI)[12] data is taken into consideration for determining the toxics or hazardous chemicals released into the water from an industry. TRI is a source about chemical toxic releases and pollution prevention events reported by industrial and federal services. This inventory acts as a central database which stores a brief information about a plethora of industries or facilities. The important thing here is to bifercate multiple regions of these facilities and analyze which facility and the region lucrative for the highest amount of toxic releases.

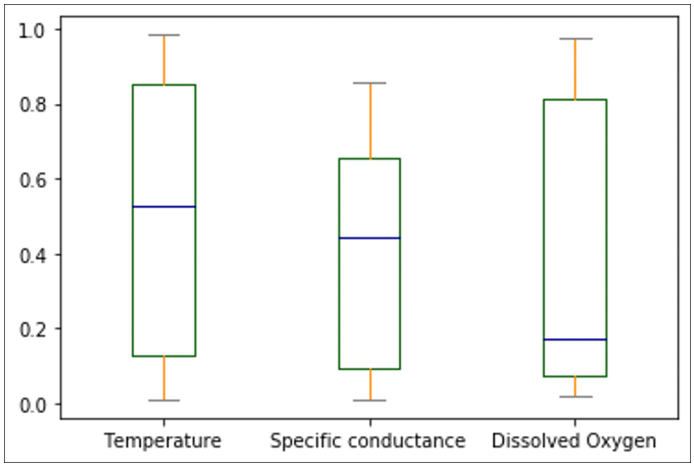

Figure 1. Distribution of key factors across the dataset

- Source: United States Toxics Release Inventory (TRI)Program
- Link : https://www.epa.gov/toxics-release-inventorytri-program
- Size of the Dataset: 5 MB
- Number of features: 11 columns
- Number of records: 50000 Tuples
- TimeLine: 2007 to 2020
- Key Features: Facility, Year, Media Type, Chemical, Releases (lbs)

B. Data Preprocessing

After the initial process of data conversion, the first thing is to perform feature engineering on the data which majorly includes handling missing features and feature encoding. Now, this dataset had a lot of Null values wherein the data for specific feature i.e. water contamination factors were missing. Hence, the null values were handled first and then the encoding was performed to convert the text data into a numeric format in order to implement machine learning algorithms on it.

C. Visualization for Factors Affecting Water Toxicity

The next important step is to visualize and get some information about the different features of the dataset. Accordingly, we performed data visualization and got some interesting results considering multiple features as factors, years and their levels.

The box plot in Figure(1) shows a distribution of key feature columns such as Temperature, Specific Conductance and Dissolved oxygen levels.

From this Figure (1), we can see that we do not have any outliers. This will ensure that the Machine model that

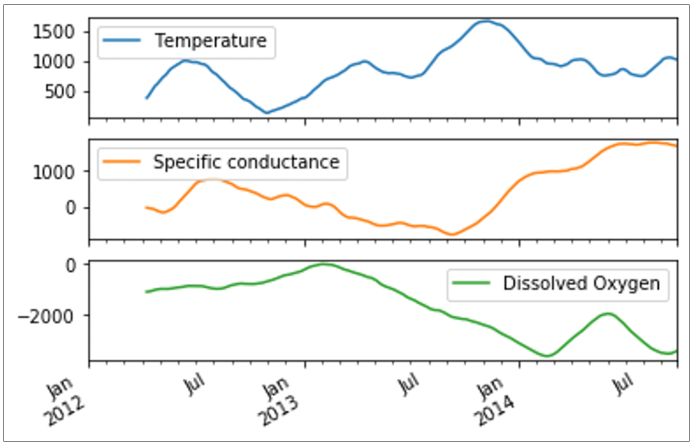

Figure 2. Time-Series plot for contamination factors over the years.

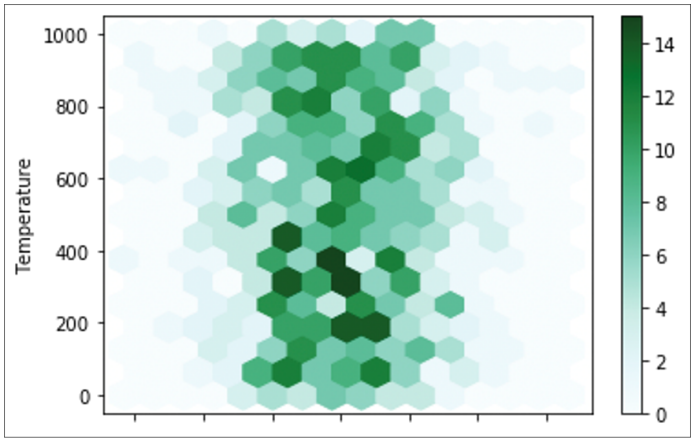

Figure 3. Visualization for Distribution of Specific Conductance of data with respect to temperature

we would apply on the data would be a great fit.We can also observe that the shape of the distribution of Temperature and Specific Conductance is positively skewed whereas the distribution shape of Dissolved Oxygen is negatively skewed. This Time-series plot shows the variation of Temperature, Specific Conductance and Dissolved Oxygen Levels[4] with Time ranging from Jan 2012 to July 2014. We use this Time Series plot to observe anomalies in Temperature, Specific conductance and Oxygen. As observed, the Specific Conductance has been increasing gradually over time and the Dissolved Oxygen has been decreasing over time. This indicates a deterioration in the quality of the water reservoirs over time.[6]

Figure(2) shows the variation of Temperature in correspondence to the PH Value across the data set with Time[1]. We noticed that at a higher temperature the acidity of the water Reservoir was extremely less. The acidity of the water reservoir increased with the Temperature.This was a key observation which is shown in 6.We also plotted the graph for the Time-series plot that shows the distribution of Temperature in water reservoirs in South Carolina across the data set with Time.We used this plot to find any anomalies of temperature at any given point by comparing it with the historical data that is obtained from the scatter plot. This was also a key observation which is shown in Figure (7), we know that conductance tend to increase as contamination increases. Figure (8) depicts the scatter plot for conductance over the past few years, the plots clearly show a linear increase in conductance indicating that the contamination is increasing over the years. This observation is shown in Figure(2).

This Time-series plot shows the variation of Temperature in correspondence to the PH Value as shown in Figure (3). The density of green color indicates the absence of Acidity in the water reservoir over a period of time.

D. Data Distribution Results

Until now, we focused on Specific Conductance as the primary factor to determine the contamination of water but after researching further we identified Dissolved Oxygen to also be a prominent factor to predict the contamination.

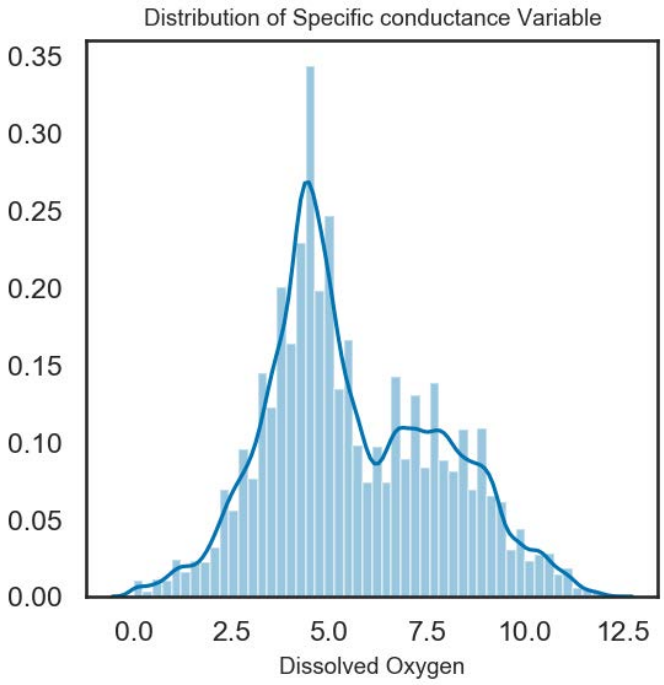

Figure 4. Visualization for Distribution of Specific Conductance.

Therefore, Specific Conductance and Dissolved Oxygen can be considered as prominent factors for measuring water's purity. The distribution of Specific Conductance of data from the water reservoirs in our dataset can be visualized using Figure (4)

We now focus on determining contamination based on the Dissolved Oxygen data. The distribution of Dissolved Oxygen[13] of data from the water reservoirs in our dataset can be visualized using Figure (5)

E. Data Imbalance and Sampling Techniques

For this approach, one hot encoding was performed in order to classify the contaminated water with the non-contaminated

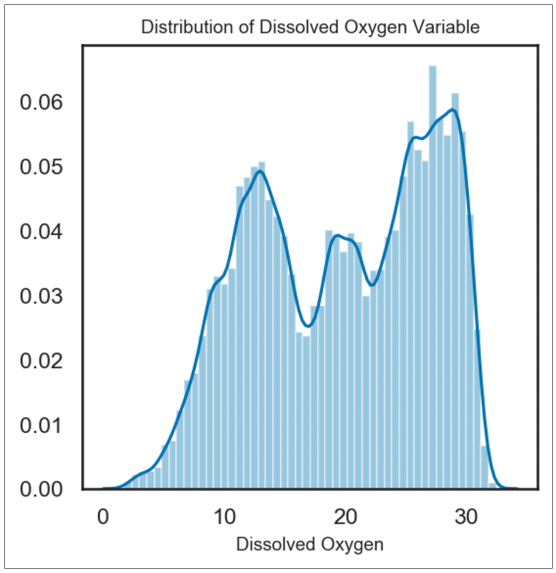

Figure 5. Visualization for Distribution of Dissolved Oxygen.

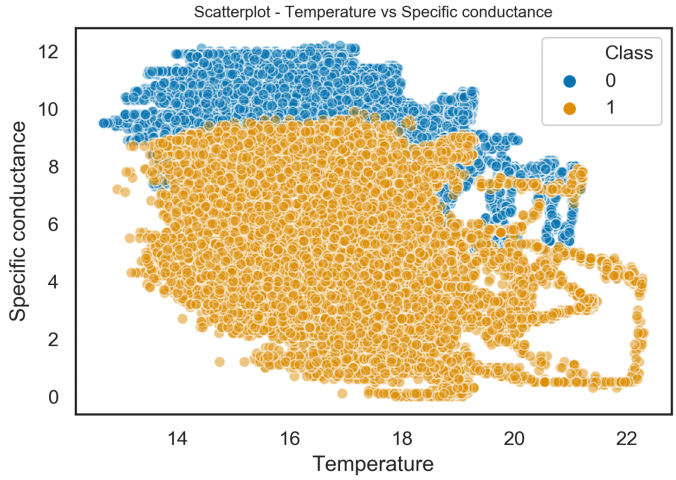

Figure 6. Visualization for temperature vs specific conductance before under-sampling

one. The aim here was to overcome the problem of imbalance and converting the data into numeric format in order to implement the machine learning algorithms. Figure (6) shows the visualization before sampling of the specific conductance.

1) Need for Sampling:

The Class feature is the predicted class, which is a binary Classification that symbolizes 0 as a contaminated water reservoir while 1 is categorized as an uncontaminated water reservoir. It can be seen that we have a lot of parity in the number of data rows among the two decision classes. The number of Class 1(uncontaminated water) is way higher than the number of Class 0(uncontaminated water reservoirs). It signifies that we have a lot of imbalance of the Class feature in the existing data set. In order to get a better fitted model, we need to decrease the above mentioned imbalance.In order to fix this issue there are Sampling techniques such as underSampling(where we will be reducing the data samples with Class 0) and OverSampling(where we will be increasing the data samples with Class 1). These methods have been utilized to fix the above mentioned problem of under-representation of Class 0 in the Class variable.

2) Applying Sampling Process:

In order to tackle the problem of imbalance in our dataset ,undersampling methodology was first applied in which the data with more number of samples was reduced to match the number of samples of the minority class and thus overcame the imbalance problem in the dataset. The IMBLEARN package's Random Undersampling method was used in order to perform under sampling of data.
We also applied oversampling to tackle the problem of imbalance in our dataset. Oversampling is the method of sampling that copies the existing observations of type Class 0 and replicates it in the data set until the number of observations of Class 0 is comparable with the number of observations of Class 1 type. For oversampling, we used the SMOTe(Synthetic Minority Over-sampling Technique): This is a technique based on nearest neighbors of Class 0. It synthesizes new Class 0 instances between existing Class 0 instances by using linear interpolation. We used the SMOTE library from the "imblearn" package for oversampling methods.
After reshuffling, we concatenated the data samples using the pandas library.After performing the undersampling and oversampling methodologies on our dataset, we estimated the change in the outcomes by measuring the accuracy of predictions and the F1 score as shown in the Table 2 and Table 3.

3) Data after Sampling Process:

To apply the machine learning models efficiently we have applied under sampling along with over sampling. The below diagram shows the scatter plot of under-sampled data with respect to class 0 (Contaminated) and class 1 (non-contaminated) data. These are the dataset representations after performing sampling techniques shown in Figure (7)

After performing the sampling methodology on our dataset, we estimated the change in the outcomes by measuring the accuracy of predictions and the F1 score for different machine learning models. These details are discussed in the next section of this report

F. Co-relation between Industry release and toxicity of Water

The Line graph in Figure (8) gives a very clear idea about the different regions where multiple facilities are located and the total emissions / releases in lbs of chemicals emitted by them. Now, the important factor to consider here is that by

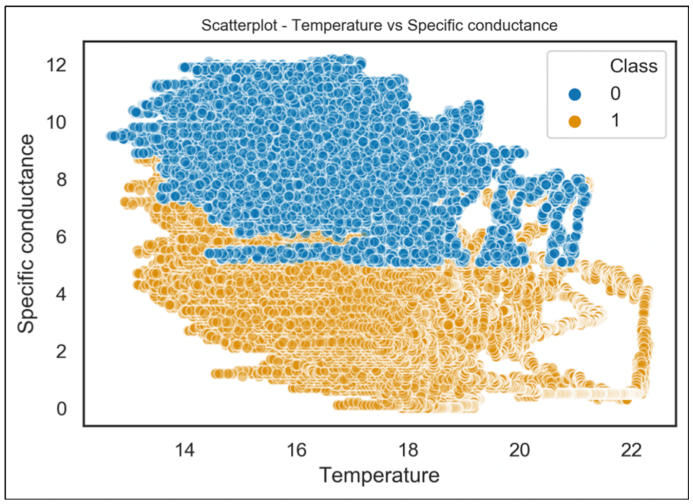

Figure 7. Visualization for temperature vs specific conductance after under-sampling.

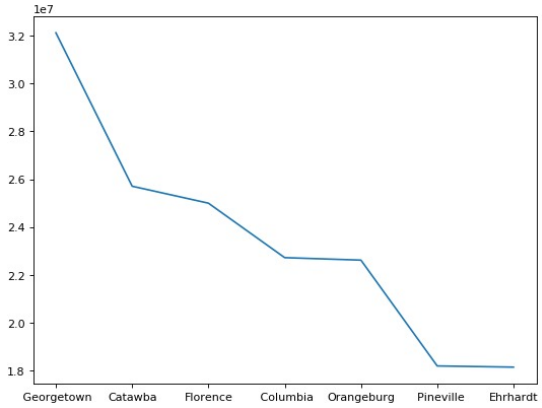

Figure 8. Visualization of Regions with respect to the releases in lbs.

viewing Figure (8) we could easily find out that the most lucrative region which has the most chemical releases out of all.

Furthermore, the bar-graph visualization in 9 fleshes out a relationship between the facilities with respect to the previous graph which facility has the most releases and the region of that specific facility. This is an important step which helps us finding the top 7 facilities with respect to their regions which is mapped with the previous data set that includes the factors influencing water contamination. In this way we understand the region-wise contamination of water in South Carolina.

The United States Geological Survey data gives information about Temperature, Specific Conductance, Dissolved Oxygen, PH Value, Precipitation, Gage, height, Timestamp of water and United States Toxics Release Inventory (TRI)[12] Program provides key insights about the release of chemicals in water with respect to each individual industry over the years in the state of South Carolina. Applying Decision Tree model, we predict the contamination of water from years 2007 to 2020. Also, when the TRI data was evaluated, we could

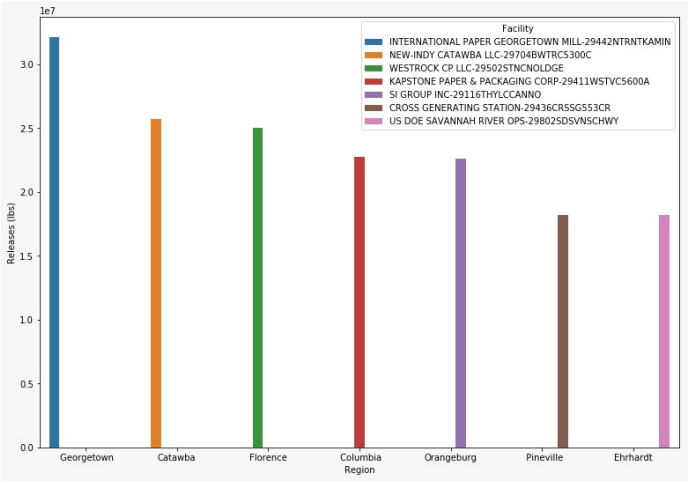

Figure 9. Visualization for regions and facilities with respect to chemical releases

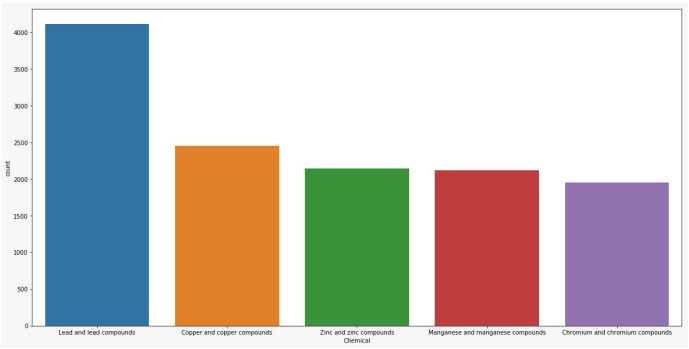

Figure 10. Distribution of Chemical counts

clearly identify a correlation between the contamination data and the release from industries. The analysis indicated an exponential rise in contamination of water with the increase in release of chemicals, which was an obvious finding. But an interesting point identified here analysis was that the major contributor to the water contamination is ”Lead and Lead Compounds”. The amount of lead dissolved in water is directly proportional to the specific conductance of water. Hence we can use this co-relation to identify the industries releasing a large amount of Lead and Lead Compounds to be the one’s majorly contributing towards the contamination of water reservoirs making the water unsafe for drinking purposes.

Below are the algorithmic Steps to determine co-relation between Industrial Chemical release and Water Contamination [4][1][5]:

1) The TRI data was grouped industry wise and the respective release count in lbs was identified in the state of South Carolina.
2) To identify the missing regions in the TRI data the information scraping was done to obtain the region data.
3) Generated effective visuals to indicate the release over the years from top 7 facilities/industries with respect to their region/neighborhood.

4) Visualized the distribution of chemicals from each industry and identified the major contributor to the toxics in water Reservoirs to be Lead and Lead Compounds.
5) Identified the top 7 Neighborhoods/Regions contributing to water Contamination, with Georgetown region being in the topmost contributor.
6) Mapped region-wise release data with the water contamination data to verify the correlation

III. Experimental Results

Model	Recall	Precision	F1	Accuracy
Logistic Regression	0.9602	0.9490	0.9546	92.4%
G Naive Bayes	0.9458	0.9577	0.9517	92.11%
Decision Tree	0.9538	0.9496	0.9517	92.05%
K-Nearest Neighbour	0.9572	0.9467	0.9519	92.06%
Random Forest	0.9597	0.9451	0.9524	92.11%

Table I
Results for UnSampled Data

Model	Recall	Precision	F1	Accuracy
Logistic Regression	0.8781	0.9288	0.9027	90.45%
G Naive Bayes	0.8649	0.9353	0.8987	90.16%
Decision Tree	0.8574	0.9320	0.8931	89.65%
K-Nearest Neighbour	0.8750	0.9271	0.9003	90.22%
Random Forest	0.8757	0.9154	0.8951	89.64%

Table II
Results for UnderSampling

Model	Recall	Precision	F1	Accuracy
Logistic Regression	0.8826	0.9238	0.9027	90.49%
G Naive Bayes	0.8658	0.9338	0.8985	90.22%
Decision Tree	**0.9089**	**0.9497**	**0.9289**	**93.04%**
K-Nearest Neighbour	0.9086	0.9281	0.9182	91.91%
Random Forest	0.9117	0.9437	0.9274	92.86%

Table III
Results for After OverSampling

The results of experimenting with attributes individually were encouraging and paved the way to an idea of combining two factors together. In this experiment we combined Conductance and Dissolved Oxygen results. Less accuracy was expected for this approach, but we obtained interesting results when both the attributes were combined to predict if the water is contaminated or not and resulted in an accuracy of **93.04%**. The results for our experiments can be summarized as follows: **Table I** represents the experiment for unsampled data considering the attributes, Specific Conductance and Dissolved Oxygen together. It can be observed that we achieved an highest accuracy of 92.4% for Logistic Regression Model.

Table II represents the experiment for under-sampled data considering the attributes Specific Conductance and Dissolved Oxygen together. It can be observed that we could carry out highest accuracy of 90.4% for Logistic Regression Model. For this the data set was decomposed into subsets as mentioned earlier. This is because the decision tree has several advantages over neural networks.

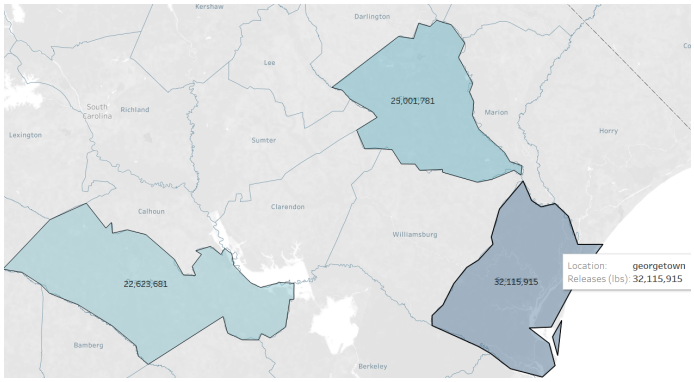

Figure 11. Visualization of most prominent water contamination regions w.r.t chemical releases in South Carolina

1) For policy maker, IDTL model is easy to understand generated tree structure.
2) IDTL model training process is faster than ANN model, and is always convergent.
3) Knowledge of IDTL model can help us choose parameters and assess the dependencies between related attributes.

Table III represents the experiment for over-sampled data considering the attributes Specific Conductance and Dissolved Oxygen together. It can be observed that we obtained a highest accuracy of **93.04%** for Decision Tree Model. Based on the above experiments we can conclude that the Decision Tree model works best with the over-sampled data taking into consideration, Specific Conductance and Dissolved Oxygen attributes yielding the highest accuracy of 93.04% with respect to all experiments.

Moreover, Figure (11) above represents a visualization of regional map of South Carolina in the United States which represents most lucrative regions with respect to the highest releases of chemicals for which the consequence is that the quality of water is affected most in those regions resulting in contamination of water.

IV. Limitations and Future Work

One of the major limitations we faced was with respect to the data source. We were unable to obtain the MERRA Satellite data since viewing of the data required national security clearance due to the sensitive nature of the dataset. Hence, we had to settle for the USGS(United States Geological Survey) dataset which was not as detailed as the MERRA satellite dataset. The data that was collected from USGS was unstructured and not easily readable through any scripts. Hence, we had to manually interpret the attributes such as PH Value, Specific conductance etc. from the .txt files. MERRA Satellite dataset link : https://gmao.gsfc.nasa.gov/reanalysis/MERRA/

Another feature we plan to implement is the Normalized Mutual Information (NMI) is a normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation). In this function, mutual information is normalized by some generalized mean

of Class 0 (contaminated data) and Class 1 (non-contaminated data), defined by the average method. We will be integrating this score along with the Precision, Recall and F1 score into our machine learning models as a part of the future work we have planned.[14] An artificial neural network learning algorithm is a computational learning system that uses a network of functions to understand and translate a data input of one form into a desired output, usually in another form. This would provide more accurate results than the machine learning models that we have applied. We intend to apply neural networks as part of our machine learning algorithms in order to extract more information and device correlation about the factors that influence water contamination as a part of the future work on this project.[15]

V. Conclusion

Specific Conductance and Dissolved oxygen can be considered as two primary factors for determining the toxicity of water. We use various machine learning algorithms to classify the water reservoir data into pure or contaminated water and the Decision Tree Model comes up with an highest accuracy of 93.04%. This Classification paves a way for determining the correlation between The United States Geological Survey data and Toxics Release Inventory (TRI) data indicating that there is a direct relationship between the water reservoir contamination in the state of South Carolina and the Chemical releases of Facilities in South Carolina. Specific Conductance of water increases as the Lead content in water increases and the Dissolved oxygen tends to decrease. The higher value of Specific Conductance and the lower value of Dissolved oxygen are indicators of water being Contaminated. Lead and Lead Compounds are identified to be the major source of contamination and the Facilities in Georgetown are the topmost contributors of toxic chemicals.[7] Based on our Analysis, we can identify and relate the Chemical releases in pounds from various industries in South Carolina over the years which helps in visualizing the top contributors for water reservoir contamination.

VI. Appendix

Below is the link to our code:

Click on ***GitHub***

or use URL:

https://github.com/Srivathsan2010/Mining_Massive_Data

References

[1] Charles Dow and Robert Zampella. "Specific Conductance and pH as Indicators of Watershed Disturbance in Streams of the New Jersey Pinelands, USA". In: *Environmental Management* 26 (Jan. 2000), pp. 437–445. DOI: 10.1007/s002670010101.

[2] L. Ding et al. "Data-gov Wiki: Towards Linking Government Data". In: (2010).

[3] Hao Liao and Wen Sun. "Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method". In: *Procedia Environmental Sciences* 2 (Dec. 2010), pp. 970–979. DOI: 10.1016/j.proenv.2010.10.109.

[4] G. Tan et al. "Prediction of water quality time series data based on least squares support vector machine". In: *Procedia Engineering* 31 (2012), pp. 1194–1199.

[5] Shuangyin Liu et al. "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction". English. In: *Mathematical and Computer Modelling* 58.3-4 (2013), pp. 458–465. DOI: 10.1016/j.mcm.2011.11.021.

[6] Toochukwu Chibueze Ogwueleka. "Use of multivariate statistical techniques for the evaluation of temporal and spatial variations in water quality of the Kaduna River, Nigeria". In: *Environmental monitoring and assessment* 187.3 (2015), p. 137.

[7] Dandan Zhao, Yang Yu, and J Paul Chen. "Treatment of lead contaminated water by a PVDF membrane that is modified by zirconium, phosphate and PVA". In: *Water research* 101 (2016), pp. 564–573.

[8] Alex Chojnacki et al. "A Data Science Approach to Understanding Residential Water Contamination in Flint". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Halifax, NS, Canada: Association for Computing Machinery, 2017, pp. 1407–1416. ISBN: 9781450348874. DOI: 10.1145/3097983.3098078. URL: https://doi.org/10.1145/3097983.3098078.

[9] Alex Chojnacki et al. "A data science approach to understanding residential water contamination in flint". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 1407–1416.

[10] Artur Jachimowski. "Factors affecting water quality in a water supply network". In: *Journal of Ecological Engineering* 18.4 (2017).

[11] Artur Jachimowski. "Factors affecting water quality in a water supply network". In: *Journal of Ecological Engineering* 18.4 (2017).

[12] S. D. Gaona. "The Utility of the Toxic Release Inventory (TRI) in Tracking Implementation and Environmental Impact of Industrial Green Chemistry Practices in the United States". In: (2018).

[13] Craig J Brown et al. "Factors affecting the occurrence of lead and manganese in untreated drinking water from Atlantic and Gulf Coastal Plain aquifers, eastern United States—Dissolved oxygen and pH framework for evaluating risk of elevated concentrations". In: *Applied Geochemistry* 101 (2019), pp. 88–102.

[14] Ping Liu et al. "Analysis and prediction of water quality using LSTM deep neural networks in IoT environment". In: *Sustainability* 11.7 (2019), p. 2058.

[15] Eric S McLamore et al. “SNAPS: Sensor Analytics Point Solutions for detection and decision support systems”. In: *Sensors* 19.22 (2019), p. 4935.