CS6316 Project Check Point: Improving Virginia Resident's Water Quality

Hanyu Li
hl7tv@virginia.edu

Monique Mezher
mmm5fd@virginia.edu

Parima Sahbai
ps4ww@virginia.edu

I. Abstract

Our project focuses on the monitoring and the analysis of the water quality of Shenandoah river in located in Virginia. Having heard of many previous cases of water quality problems throughout the United States, we decided that independently predicting water quality based on available data sets online would be beneficial to many Virginia residents. Although water standards and quality measures area already in place in much of the United States, many of these regulations have become about politics and an independent investigation would be helpful to determine a threshold for safe water based on raw data. Our goal was to identify trends and to pinpoint problem areas based on the original data set provided by "The Friends of the Shenandoah River" which is a volunteer, non-profit, and scientific organization. This data set contained various quantified measures of the water attributes at several sites along the river. Using raw data about the water, such as its pH, its nitrate and phosphate concentration, and its turbidity, we hope to construct a model enabling Virginia residents close to the water have a better idea about the health of their river ecosystem. This paper will focus on our original motivations for pursuing this experiment, our current methodology, initial discoveries and planned future work.

II. Motivation

Water is an essential part of the life and the well-being of all citizens. Recently there has been many reports of problems with water quality and safety throughout the United States. A prime example of this issue is the extreme case of Flint, Michigan where led was found in the water, negatively affecting the health of its residents.

For our project, we are going to focus on identifying trends and problem areas about the water quality of the Shenandoah River throughout Virginia. Although the river water does go through filtering and purification before landing in Virginias tap, it is important to identify the starting composition of water in order to properly treat it. Furthermore, animals and plants native to Virginia could also benefit by being able to deter them from the river during unclean seasons. Virginia residents expect healthy drinking water and and we want to make sure that residents are provided with safe-to-drink tap water, and are aware of their ecosystem.

Since our project deals with a topic that is critical to everyday life of Virginia residents, the results found could be applied to provide more efficient water purification and guaranteed quality. Patterns found after analysis of water quality throughout time could help predict when certain chemicals or bacteria may bloom, and therefore determine when water quality is at its lowest. The results will also be a beneficial resource to those who reside along the river, particularly if they use the water for watering plants, septic systems or emergency water supply.

III. Method

1) Data pre-processing: The original data set was divided into several .xls files of different according to date, so we combined them into one file and converted it into a .csv file using scripts and Python.
2) Data cleaning: The data set had a number of missing values for several different features that could not be simply be disregarded. In order to fix this issue we decided to impute the empty values with mean values from our data set. For some of the features, the ratio of missing values was over 90. This resulted in the features being deleted since imputing values into such features seemed meaningless.
3) Label creation: The original data set simply recorded each site's water condition and along with lab comments. In order to apply the data set to machine learning models, we needed to create a label column to use as our "Y" to reflect the samples' water quality. After some background research, we decided to evaluate the water quality by a new parameter "Health Score" which was calculated using existing features.
4) Research towards Determining a new parameter, "Health Score": Through research we were able to future out and make use of the raw data that we were provided to create this feature. According to our research, we were able to understand our raw features more which we will further explain.
    - Turbidity: a measure often used to determine the amount of pollution in a body of water. The higher the Turbidity, the higher the likely of pollution: 1.00 g/m3 = 1.00 mg/L = 1.00 ppm.
    - Nitrate PPM: Nitrate is naturally present in water and soil, it is a dissolved form of Nitrogen. The normal amount of Nitrate are usually below 1.00 mg/L. Concentrations of Nitrate over 10.00 mg/L

can have negative effects on the safe use of water and its aquatic environment.
    - Ortho-Phos PPM: A measure of reactive Phosphate which is the hydrolysis or the chemical breakdown of Phosphate which naturally exists in bodies of water. In general, concentrations of Ortho-Phos over 0.05 mg/L will likely have an impact while concentrations greater than 0.10 mg/L will certainly have impact on a river. The most ideal and healthy river would have a concentration of 0.00 mg/L.
    - Amm PPM: Ammonia is another form of Nitrogen which naturally exists in bodies of water. Elevated concentrations of Ammonia are directly toxic to an aquatic environment. A concentration of 1.90 mg/L is considered toxic, and the most ideal concentration is 0.00 mg/L.
    - Lab pH: pH levels are measures of the acidity, specifically hydrogen ion concentrations in water. The optimum pH for river water is a pH of 7.4.
5) Model training and comparisons: The probability of randomly picking a single model, and getting the best results through it is incredibly low. In order to get the best fit, we need to test multiple different models and decide based on several different measures such as RMSE and MAE. For this project, we chose three different models to work with: Random Forest Regression, Linear Regression and Stochastic Gradient Descent (SGD) Regression.

IV. Preliminary Experiments

We tried three different models to find out which performs better given our data set. These models include Random Forest Regression, Linear Regression and Stochastic Gradient Descent Regression. Each of the team members trained a different model using the same data set that had been cleaned beforehand. In case of RMSE and MAE, SGD Regression seems to be the best one among the three models we've tried, while Random Forest Regression has an advantage over other methods with its ability to deal with more categorical features.

V. Next Steps

1) Tune-Model experiments: Once we have chosen a model after careful consideration of its abilities and its fit, the next step should be focusing on tuning this model to better its performance as much as possible to achieve a more accurate and reliable predictor.
2) Feature selection and combination: Currently we have around eight features, which does not constitute a large number for making predictions as it is. However, we can still try different feature combinations to see if better predictions could be made and/or predictions stay as they are but with less features and less unnecessary data overhead which may slow down the model with nonessential features.
3) Match real locations on the river with the site ID features: As a part of our presentation, we want to visualize our analysis on a map to give a more straightforward and visual sense of water quality trends along the Shenandoah River. In order to make said map, we need to find the real locations of the "Site-ID" feature in our data set which is designated to determining the location where the data was gathered from. After gathering these locations, we need to come up with the map in order for us to use it in the final video project.

VI. Contributions

1) Hanyu Li: Did the data cleaning part. Applied the method of calculating "Health Score" and added the Y column to the cleaned dataset. Tried SGD Regression and measured its performance on the prepared dataset. Finished some of the real-location-finding work and recorded the sites' information in a table for further reference. Wrote the preliminary version of this check point document.
2) Parima Sahbai: Ran Random Forest Regression on the data set and edited/wrote parts of the check point document. Will be working on the video/editing the final results.
3) Monique Mezher: Did data aggregation from the raw data on the source website into readable csv format. Tried linear regession and measured its performance on the prepared dataset. Created the document and wrote the preliminary version of the checkpoint document.

References