Final Report:

Georgia Water Quality Health Analysis

Problem Statement

Bodies of water, such as streams and lakes, are dynamic features of the natural world. The composition of water bodies change throughout the year as the seasons change. Sources of water such as inlets and storm water runoffs impact the quality of water in these reservoirs. Points of pollution from factories and farms push in organic material such as nitrogen and phosphorus that can raise their respective concentrations above a safe threshold, causing damage to the native wildlife that live in and around the body of water. Is it possible to model these reservoirs, based on their characteristics, to predict which components greatly impact the standard water quality indicators the most?

By researching the United States Geological Survey (USGS) and Environmental Protection Agency (EPA) national database on water quality data, I created a tool that can predict water quality readings by reducing the amount of sampling needed to draw conclusions. Many techniques were used to establish this model through Supervised and Unsupervised Machine Learning, Geostatistics, and Data Visualization.

The original dataset was reduced from over 100 features to 12, and the Random Forest Classification model was able to achieve an average precision of 0.95 for 91 sites. This process can be repeated for any state, reservation, or hydrologic region in the USGS database for water quality data processing improvements.

Data Wrangling

The raw data set was composed of 78 columns and 105177 rows representing 91 sampling sites. A pivot table was required to convert the columns with technical features as columns instead of rows. By indexing by location instead of by type of entry, the data set was reduced to 91 rows and 12 columns. Null values were filled with the median as majority of each feature was found to be not normally distributed.

Exploratory Data Analysis

Water quality thresholds[1] are dictated by the EPA as follows:

• pH:	6.5-9
• E. coli:	< 410 (cfu/100ml)
• Dissolved Oxygen (DO):	> 3.0
• Turbidity (NTU):	< 2.34
• Total Nitrogen:	< 10 (mg/L)
• Total Phosphorus:	< 21.88 (ug/L)
• Chlorophyll a:	< 30 (ug/L)

There are quite a few more parameters that determine water quality, but these are considered primary indicators of water quality by the EPA. In fact, for the purposes of this

[1] https://www.epa.gov/wqc/national-recommended-water-quality-criteria-aquatic-life-criteria-table#table

modeling study, only E. coli is specifically examined as an indicator of poor water quality. The other parameters are seen as independent variables.

I looked at correlations between every column and my target parameter, and I found some strong correlations which I then graphed and plotted (Figure 1 and 2). The strong correlations had to have an absolute value greater than 0.5.

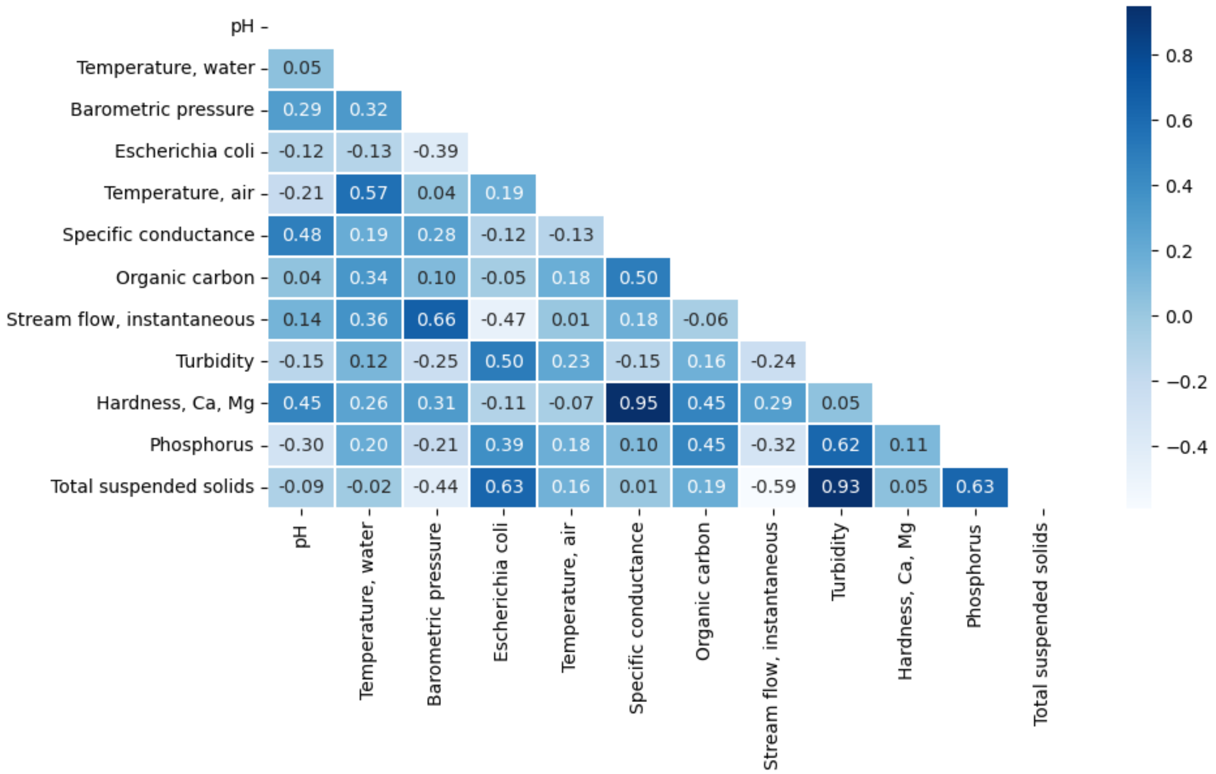


Figure 1: Heatmap Pyramid of all Variables. Correlation strength is shown in each block.

After examining correlations between the parameters in the data set, the only strong correlations for E. coli were total solids (0.63) and turbidity (0.5). Inverse correlations were found to be formed with pressure (-0.47) and stream flow (-0.47).

Intuitively, this makes sense as higher E. coli counts are usually found to be caused by macroscopic objects and coliforms that also add to the non-dissolved solids count and the turbidity opaqueness. Previous research has shown that high hydrostatic pressure filtering is linked to bacterial coliform death. Faster currents do not allow coliforms to pool together in a cloud in the water causing adverse conditions for coliform growth.

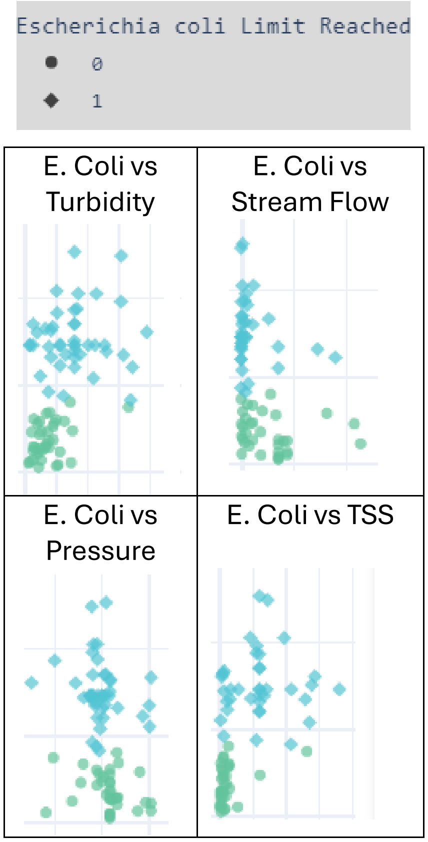


Fig. 2. These scatter plots show the relationship between E. Coli. and its strongest correlative parameters. The plots are also flagged with symbols to show which sampling was over the threshold value for E. Coli. deeming it unhealthy water sources.

In-Depth Analysis

Now that I considered the relationship between different parameters, I need to determine the number of components would be best to look at for the model. After this, I can build and compare different models to see which predict the test data the best. With that, I will note the accuracy, recall, and f1-score using a confusion matrix.

PCA (explain x vs y)

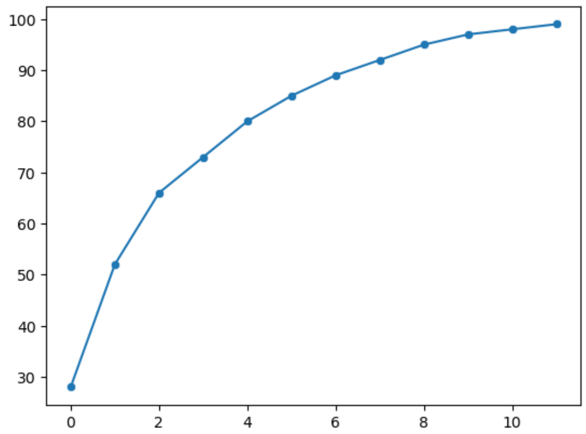


Fig. 3. The number of features is listed on the x axis and the percentage of the explained variance is listed on the y axis. Seven features are required to explain 90% of the variance in the dataset. This means that there is no need for a reduction in features as that requires mostly all of the components.

Model Selection

For each site, I tested 8 models to experiment with to find baseline models and compare performance with each other: Logistic Regression, Support Vector Classification, K Neighbors Classifier, Gaussian Naïve Bayes, Ridge Classifier, Random Forest Classifier, Ada Boost Classifier, and Gradient Boosting Classifier.

All models have moderately high to extremely high precision scores even before tuning the parameters of the models. XGB performed the best with a 0.95 precision score and there was a three-way tie between Logistic regression, Ridge classifier, and Random Forest classifier. XGB is interesting because of its unique ability to handle null values in its classification, while random forest already has the tuned hyperparameters of fundamental modeling schemes such as Decision Tree classifier and Bagging classifier.

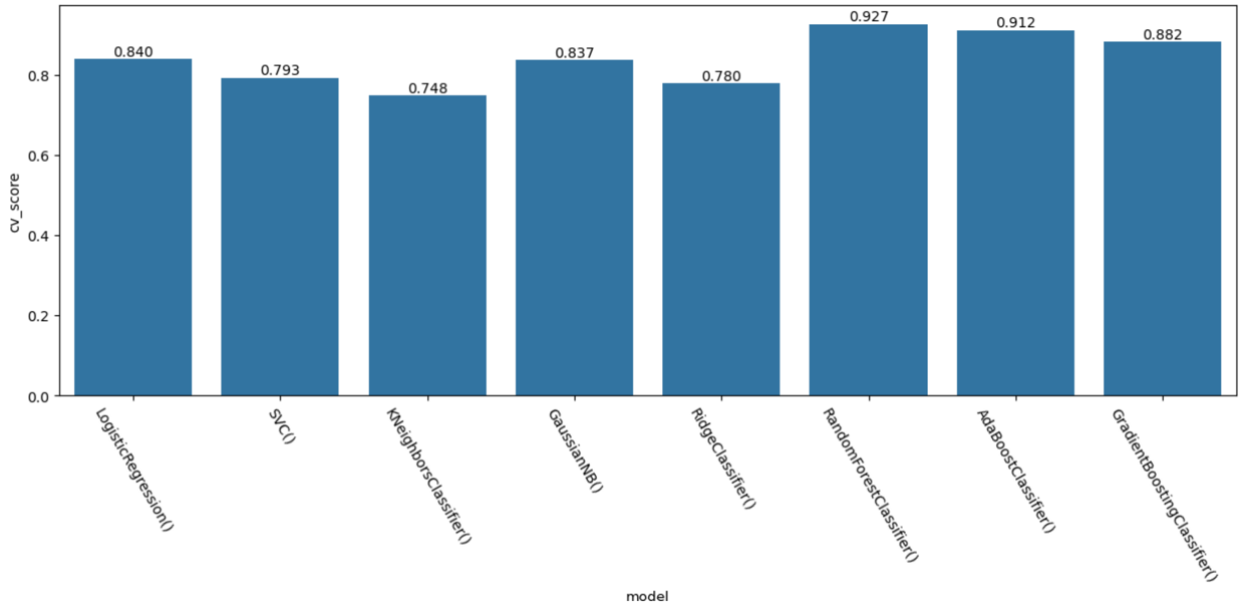
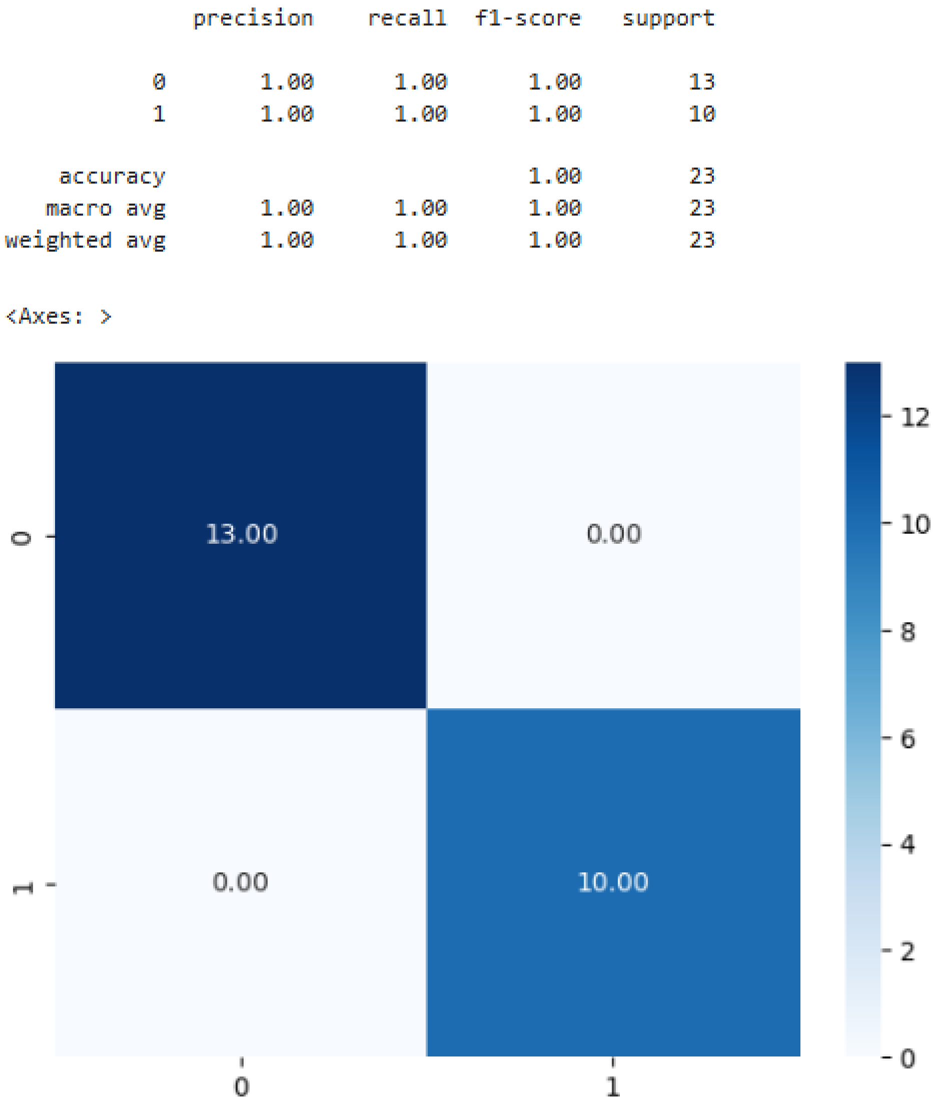


Figure 4: Model Comparison

Hyperparameter tuning was conducted by means of a repeated stratified K fold. Parameters were explicitly given and looped through five splits, repeated twice for the two best models (XGB and Random Forest), to find the best parameters. After cross validation was conducted, both XGB and Random Forest had the same precision score of 0.95614, increasing Random Forest by 0.4 points and slightly increasing XGB by 0.3 points.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        13
           1       1.00      1.00      1.00        10

    accuracy                           1.00        23
   macro avg       1.00      1.00      1.00        23
weighted avg       1.00      1.00      1.00        23

<Axes: >
```

Figure 5: Confusion Matrix

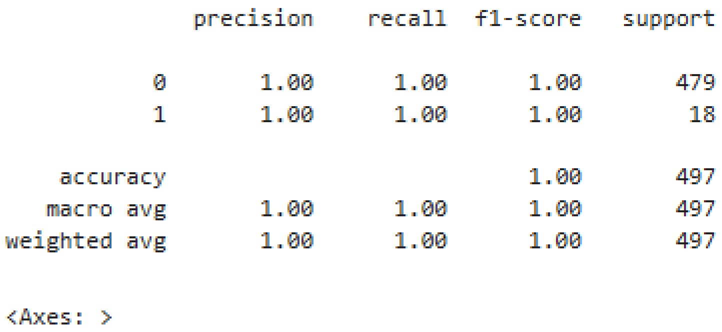
Afterwards, a voting classifier combined the two models together, taking in their results, and forecasts a class based on the class with the highest likelihood of becoming the output. The accuracy of this combined voting classifier was determined to be higher than the individual models with a precision score of 0.961.

Validation Test

I pulled a random assortment of sites across the United States from the period, 07/2023-08/2023, in order to validate the generated model.

After downloading an approximately 100 MB dataset of all sites containing E. Coli. samples, the dataset was further reduced to contain only the E. Coli samples, which resulted in 496 sites. I created the E. Coli. threshold flag again with a peculiar distribution. Almost all of the them except for 18 sites were within healthy guidelines. This differed from the training dataset that was split evenly.

I ran the models, XGBoost and RandomForest, on this dataset and found the accuracy scores to be perfect or near perfect. This performs better on the validation set rather then the test set.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       479
           1       1.00      1.00      1.00        18

    accuracy                           1.00       497
   macro avg       1.00      1.00      1.00       497
weighted avg       1.00      1.00      1.00       497

<Axes: >
```

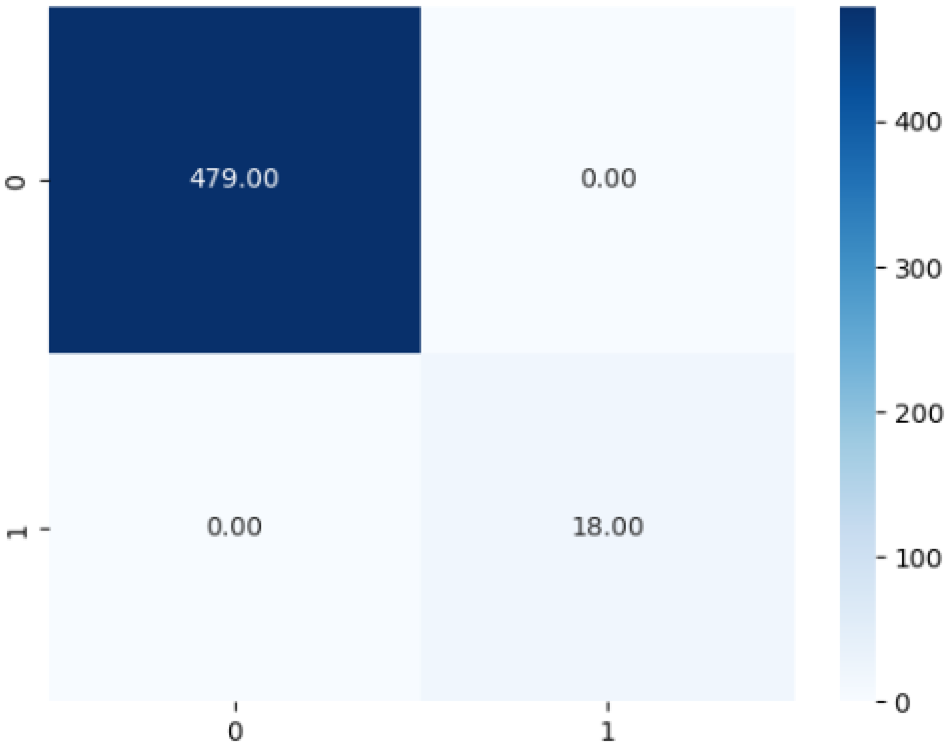


Figure 6: Confusion Matrix for Validation Set: Demonstrating no false positives/negatives

Conclusion

After composing models based on this data set, it was found that the mean coli form counts for E. Coli. is more than twice as high as the threshold value (997 to 410). Regardless, the number of water sources that are healthy is about even with the unhealth water sources (47 vs 44). The strongest correlation is with E. Coli. is TSS with the highest correlation of TSS being turbidity. This shows a three-way relationship between the parameters. Random Forest and XGB worked the best to train the model over all other baseline models. Finally, the aggregate method of using a voting classifier resulted in the highest precision score (96%).

Future Work

A more sophisticated approach would be to take the failures of E. Coli. and expand it to the failures of all other variables. This would almost double the number of features but allow insight into the correlation between failures as well as their parent parameters.

Certain sites had much different quantities than other sites. This is due to the large variability that streams and rivers provide in the values of their parameters. As lakes and reservoirs may have low, if any, stream flow, rivers and streams could move rapidly as the size of these water bodies increase.

The use of the aggregate voting classifier allows easy model production of each water quality parameter. The EPA legislates for an overarching model that combines all submodels together.

The results of this study are very compatible with analysis of water bodies from all over the United States as any state, native american reservation, or hydrologic region can be examined using this model.

Separate models can be formed based on time of day, season, and year. The data can also be taken as is from the raw dataset instead of aggregated together at the mean like in this modeling study. This would require many more models to be created and combined for comprehensive analysis.