Towards Real-Time Water Quality Assessment With Bayesian Networks

*Joshua Thompson, *Brock Weekley, *John Wisnewski
University of Missouri

Abstract— **Rapid urbanization, agricultural development, and industrial growth have all increased the demand for clean water worldwide. Likewise, the need for water quality monitoring systems is on the rise to ensure the efficient and ethical consumption of this resource. However, water quality assessment is often a lengthy process that requires sampling and post-analysis by many experts. Since many settings need high-quality real-time data, human effort alone is unmanageable. In this report, we explored applying Bayesian network machine learning to automate the quality analysis process. Specifically, we designed a network that leverages 7 physical-chemical measurements of water samples and learned to predict the quality of water from Missouri. We compare this network to the results of another machine learning algorithm, the supervised SVM model. Our results show that Naive Bayes is suitable for detecting water quality, even with missing parameters, and performs similarly to the contrasted SVM model, with greater flexibility.**

I. Contributions

All group members contributed equally to all parts of this project and report.

II. Introduction

The demand for usable water is growing at a rapid rate. In the United States alone, it was reported in 2015 by the United States Geological Survey that the daily water consumption was over 322,000 billion gallons across agriculture, domestic, industry, and other uses [1]. This demand is only expected to intensify as climate scientists predict that longer and more intense droughts will strain the clean water supply lines in the coming decades [2]. However, clean water access is not just a problem for the future but one with real consequences right now as well. In 2016, 4 billion people faced freshwater scarcity for at least one month out of the year and half of one billion people face it year-round [3]. More recently, aging infrastructure also threatens to contaminate water across the United States. Lead pipes are still the main transporter of water from storage facilities to homes for many Americans (especially in the Midwest) [4]. The dangers of such decaying infrastructure are well understood and have been seen firsthand in towns such as Flint, Michigan [5,6]. Access to clean water is also jeopardized by industrial development like fracking and oil pipelines [7-9]. In all of these cases, having access to real-time water quality data is crucial to protecting local populations and making informed policy decisions. Water quality indexes are not a one size fits all solution to water quality, but the measurement of water in this way can provide a concise summary that is useful for recognizing dangers in water and areas that need improvement. Traditional water quality assessments require sampling and analysis by experts which makes a real-time system unmanageable by humans alone.

Water quality as an index is an aggregated resource measured with physical-chemical or microbiological recordings. The first water quality index was introduced in 1965 and various alternative indexes have been introduced since then that vary in regional attributes and parameters [10-13]. Some prior works have attempted to apply neural networks to water quality assessment, but found that such models often suffered from regional bias [14,15]. Additionally, since water quality data can often include missing parameters and require interpretable

* denotes equal contributions

results, neural networks have typically been found to be unsuitable for this task.

To manage the uncertainty observed in water quality data, Ilic et.al. instead leveraged a naive Bayes approach [16]. In this work, they attempted to predict the water quality score of different areas in Serbia. To do this, they used the Serbian Water Quality Index (SWQI) which relies on 10 physical-chemical and microbiological parameters. The final score is determined by a weighted sum of each of these parameters and the overall quality is based on fixed thresholds from this score. For their experiments, they utilized 68 samples collected from the Serbian Environmental Protection Agency at 5 locations across Serbia. These samples were collected in April & August from 2013 to 2019. Their network architecture followed a simple structure with 9 parameter nodes that were each classified into 5 groups before training based on parameter-specific thresholds. The target water quality was also split into 5 classes. Then, using the Netica software, they used 48 of their 68 samples to train this Bayesian network [17]. Unfortunately, using a strict evaluation, their model could only classify 51 out of the 68 samples correctly, including the training data. This was likely due to poor coverage of their overall parameter space. For example, they reported in their evaluation that they only had 4 'Excellent' water samples and appeared to have misclassified all of them. In this report, we describe our efforts to improve upon these results.

III. Problem definition

Clearly, there is a demand for a generalizable real-time water quality prediction system. To that end, we attempted to recreate the network architecture utilized in [16] and apply it to new water samples collected in Missouri. We also implement a support vector machine (SVM) to serve as a baseline and further verify the variance claims made in [14] about traditional machine learning methods for water quality data. This was done to both test the validity of a Naive Bayes net approach to water quality and to contrast it against other approaches. Missouri water data was chosen to test if the approach originally used on Serbian water would be viable in other regions, with different water quality indexes, and with different parameters available.

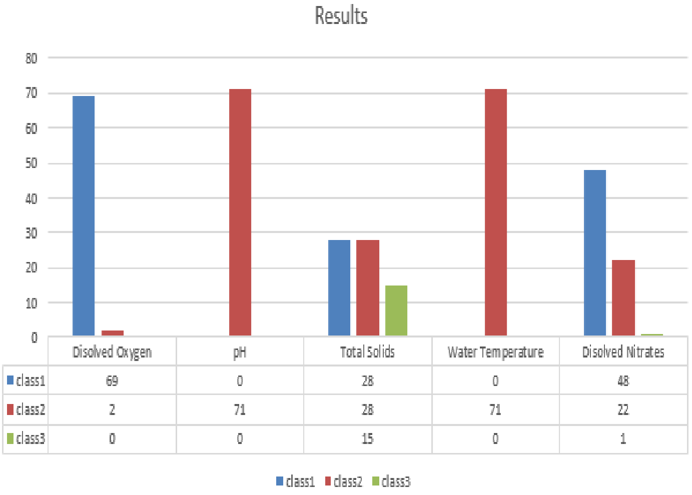
IV. Methods

Since the authors of the original proposal for Bayesian network usage in WQI did not release the data used, we are unable to directly replicate their results. Instead, we applied their methods to water samples from sites around Missouri. Specifically, we collected recordings from the USGS's report on ambient water-quality monitoring in Missouri from 1993-2017 [18]. This report provided detailed recordings of many parameters across 75 locations in Missouri from 1993 to 2017. Since we are using Washington state's water quality index, we extracted the mean recordings of coliform, ph, temperature, phosphates, suspended solids, nitrates, and dissolved oxygen from each location to get 75 samples. Next, we calculated the water quality of each sample directly using Washington's index. The various qualities for this index are split into three categories of concern: Low, Moderate, and High. These levels of concern served as our target classes to learn. Similar to [17], we also classify each of the parameters into 3 classes based on parameter-specific thresholds set by the state of Washington. Then to build our Bayesian Network, we follow a schema very similar to [17], but use the GeNIe 3.0 Academic software instead of Netica.

To compare our results with a traditional machine learning method, we also implemented an SVM. SVMs, also known as max-margin classifiers, try to pick the best linear separator by maximizing the margin between support vectors (which are the points closest to said linear separator). Since not all data is linearly separable in lower dimensions, SVMs can be further extended to higher dimensions via kernel functions. However, to avoid a model with high variance, we chose a linear SVM instead of a non-linear model. Additionally, we assign equal weight between the classification loss and the regularization term to promote a more generalizable decision boundary

V. Experimental Results

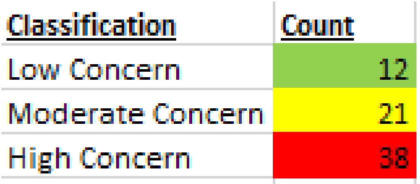
The results of the Naive Bayes Net implementation were obtained by training the data on 71 of the 77 samples. We then validated the data on the remaining samples. Figure 1 shows the distribution of each parameter into one of three classes. We then took the parameter classes for each sample to calculate their Washington Water Index score. Based on this score each sample was classified as Low, Moderate, and High concern. The calculations for the Washington Water Index and parameter classifications were both provided from the Washington State Water index. Figure 1 shows how the samples for each feature were classified.



	Disolved Oxygen	pH	Total Solids	Water Temperature	Disolved Nitrates
■ class1	69	0	28	0	48
■ class2	2	71	28	71	22
■ class3	0	0	15	0	1

Fig 1. Distribution of Feature Classifications of Original 71 Samples

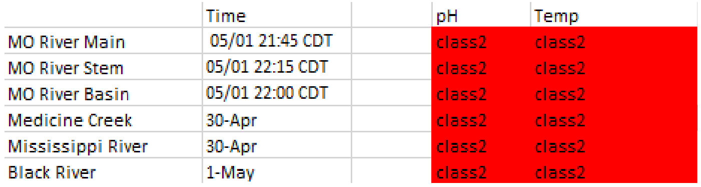
Figure 2 shows how each of the 71 samples was classified overall. More than 50% of the samples were labeled as High Concern, while less than 30% were of Moderate Concern and approximately 17% were Low Concern.



Classification	Count
Low Concern	12
Moderate Concern	21
High Concern	38

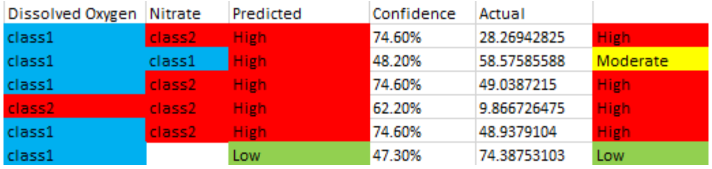
Fig 2. Distribution of Overall Classifications of Original 71 Samples

Tables 1 and 2 show how our Bayesian Network performed on the held-out samples after training. Of the six samples, our network classified five of them correctly. The second sample was predicted to be of High concern but was actually Moderate. In this case, although it was misclassified this was a conservative misclassification. That is to say, being labeled as High concern will call for more cautious treatment than necessary. This could also indicate the need for more testing on-site which would show that the sample was actually Moderate. For our application, an overcautious model would be very detrimental. Our Bayesian Network can also handle missing data. The last sample in Tables 1 & 2 had a missing value for the quantity of nitrates in the water. The classification still came out correct which demonstrates some level of adaptability for this method.



	Time		pH	Temp
MO River Main	05/01 21:45 CDT		class2	class2
MO River Stem	05/01 22:15 CDT		class2	class2
MO River Basin	05/01 22:00 CDT		class2	class2
Medicine Creek	30-Apr		class2	class2
Mississippi River	30-Apr		class2	class2
Black River	1-May		class2	class2

Table 1. Naive Bayesian Network Results on Six Recent Water Samples Pt 1/2



Dissolved Oxygen	Nitrate	Predicted	Confidence	Actual	
class1	class2	High	74.60%	28.26942825	High
class1	class1	High	48.20%	58.57585588	Moderate
class1	class2	High	74.60%	49.0387215	High
class2	class2	High	62.20%	9.866726475	High
class1	class2	High	74.60%	48.9379104	High
class1		Low	47.30%	74.38753103	Low

Table 2. Naive Bayesian Network Results on Six Recent Water Samples Pt 2/2

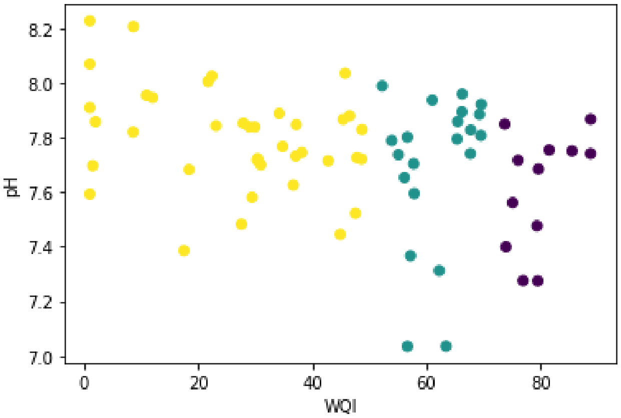
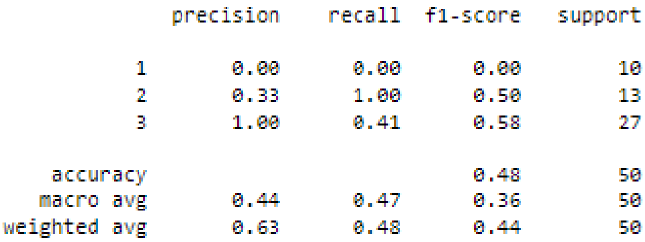


Fig 5. Results of Two Feature SVM Supervised Model

Next, we looked to compare our Bayesian network performance to an SVM. In Figure 5 we plot each sample's pH value against its overall water quality. We observe that the data is very separable when we examine only the pH and Water Quality Index Score. We note that other features compared with the Water Quality Index are also very separable. Using an SVM to model this separability differs slightly from the Naive Bayes approach since the Bayesian Network used the pH, Temperature, Dissolved Oxygen, and Nitrate levels to compute the Washington Water Index score. In contrast, the SVM uses the WQI score that is already found and uses the ranges for the WQI scores to put the samples into separate classes.. For this method 50 of the 71 samples are used for training data and 14 of the remaining 21 samples are used for validation.



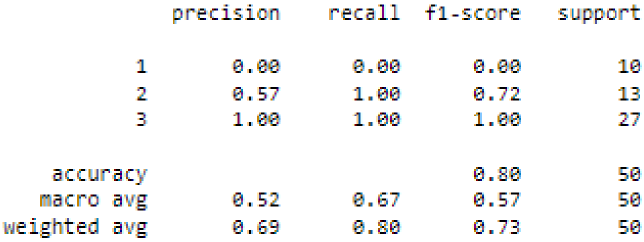
```
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        10
           2       0.33      1.00      0.50        13
           3       1.00      0.41      0.58        27

    accuracy                           0.48        50
   macro avg       0.44      0.47      0.36        50
weighted avg       0.63      0.48      0.44        50
```

Fig 6. Initial Results of SVM on Training Data

Figure 6 shows the results from the training data immediately after learning. The SVM misclassified all the points in class 1 which led the F1 score for class 1 to be zero. We then ran an optimizing function to find the best c and gamma values for this data. The c value represents how accurate we want the testing data to be and the gamma value is how much weight is applied to each sample. Given the small sample size, we found the optimal c value to be 0.1 and the optimal gamma value is 1. After these hyperparameters are applied to the SVM function we get the results seen below.



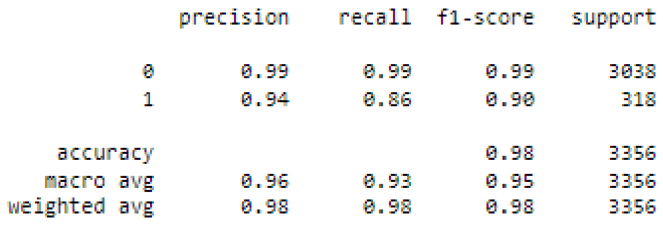
```
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        10
           2       0.57      1.00      0.72        13
           3       1.00      1.00      1.00        27

    accuracy                           0.80        50
   macro avg       0.52      0.67      0.57        50
weighted avg       0.69      0.80      0.73        50
```

Fig 7. Training data results of SVM after Optimization

Figure 7 shows improved results with a perfect f-score for class 3 and a large improvement for class 2. To see how these results generalized, we then repeated this process on the dataset we set aside for testing. Of the remaining 21 samples, we validated the SVM with 14 of them. Figure 8 shows the F1 scores before the c and gamma optimization on the validation sample set.



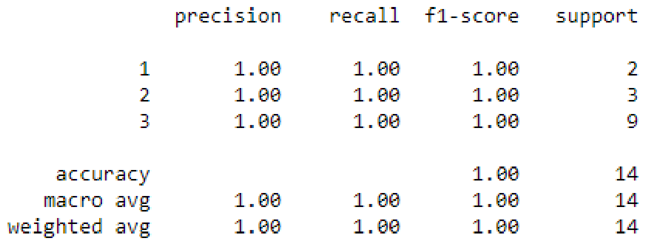
```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      3038
           1       0.94      0.86      0.90       318

    accuracy                           0.98      3356
   macro avg       0.96      0.93      0.95      3356
weighted avg       0.98      0.98      0.98      3356
```

Fig 8. Results of SVM before C and Gamma Optimization

The optimized c and gamma values for this data set are the same as above, c = 0.1 and gamma = 1. Figure 9 shows the results after the fine-tuning c and gamma.



```
              precision    recall  f1-score   support

           1       1.00      1.00      1.00         2
           2       1.00      1.00      1.00         3
           3       1.00      1.00      1.00         9

    accuracy                           1.00        14
   macro avg       1.00      1.00      1.00        14
weighted avg       1.00      1.00      1.00        14
```

Fig 9. Results of SVM after C and Gamma Optimization

After optimizing, we get perfect F1 scores which indicate perfect recall and precision meaning all validated points were identified and classified correctly. It is not likely to get a perfect F1 score every iteration but with the given data and sample size we were able to achieve this. Notably, there is an imbalance between the classes which could cause the F1 score to vary depending on which samples were used. The samples that we chose were selected at random and with so few samples that fall into class 1, it is possible for there to be zero samples that are in this class. However, it may be possible that class 1 is simply an uncommon classification for the samples in general, which would result in regular F1 imbalance.

VI. Conclusion

In this work, we explored two different methods for learning a real-time water quality assessment system. Our Bayesian Network successfully generalizes to new water samples and appears to air on the side of cautious estimates. Our SVM approach also demonstrated strong training and testing performance. Both methods of learning show very promising results. The Bayesian Network, however, appears to allow for more flexibility in required parameters and hidden variables. In future works, we would like to explore these methods with larger sample sizes. These methods provide a strong basis for understanding and testing with larger data sets and would affirm our understanding of how accurate and reliable these methods truly are. More data and testing could provide useful insight on what modifications need to be done to achieve consistent, optimal results. Overall, water quality testing automation can likely be performed with either method, but the Bayesian Network would be more efficient and flexible.

REFERENCES

1. Dieter, Cheryl A. *Water availability and use science program: Estimated use of water in the United States in 2015*. Geological Survey, 2018.
2. Fuller, Amy C., and Michael O. Harhay. "Population growth, climate change and water scarcity in the southwestern United States." *American journal of environmental sciences* 6.3 (2010): 249.
3. Mekonnen, Mesfin M., and Arjen Y. Hoekstra. "Four billion people facing severe water scarcity." *Science advances* 2.2 (2016): e1500323.

4. Brown, Mary Jean, and Stephen Margolis. "Lead in drinking water and human blood lead levels in the United States." (2012).
5. Triantafyllidou, Simoni, and Marc Edwards. "Lead (Pb) in tap water and in blood: implications for lead exposure in the United States." *Critical Reviews in Environmental Science and Technology* 42.13 (2012): 1297-1352.
6. Butler, Lindsey J., Madeleine K. Scammell, and Eugene B. Benson. "The Flint, Michigan, water crisis: A case study in regulatory failure and environmental injustice." *Environmental Justice* 9.4 (2016): 93-97.
7. Jackson, Robert B., et al. "The environmental costs and benefits of fracking." *Annual review of Environment and Resources* 39 (2014): 327-362.
8. Howarth, Robert W., Anthony Ingraffea, and Terry Engelder. "Should fracking stop?." *Nature* 477.7364 (2011): 271-275.
9. Delin, G. N. *Ground water contamination by crude oil near Bemidji, Minnesota*. US Department of the Interior, US Geological Survey, 1998.
10. Horton, Robert K. "An index number system for rating water quality." *J Water Pollut Control Fed* 37.3 (1965): 300-306.
11. Liou, Shiow-Mey, Shang-Lien Lo, and Shan-Hsien Wang. "A generalized water quality index for Taiwan." *Environmental monitoring and assessment* 96.1 (2004): 35-52.
12. Cude, Curtis G. "Oregon water quality index a tool for evaluating water quality management effectiveness 1." *JAWRA Journal of the American Water Resources Association* 37.1 (2001): 125-137.
13. Veeraswamy, G., et al. "water quality assessment in terms of water quality index in Gudur area, Nellore district, Andhra Pradesh." *Int. J. Tech. Res. Sci* 3 (2018): 1-6.
14. El Bilali, Ali, and Abdeslam Taleb. "Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment." *Journal of the Saudi Society of Agricultural Sciences* 19.7 (2020): 439-451.
15. Emamgholizadeh, Samad, et al. "Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models." *International Journal of Environmental Science and Technology* 11.3 (2014): 645-656.
16. Ilić, M., Z. Srdjević, and B. Srdjević. "Water quality prediction based on Naïve Bayes algorithm." *Water Science and T*echnology (2022).
17. Netica 2020 Norsys Software Corporation. Available from: https://www.norsys.com/download.html (accessed May 2022).
18. Richards, Joseph M., and Miya N. Barr. *General water-quality conditions, long-term trends, and network analysis at selected sites within the Ambient Water-Quality Monitoring Network in Missouri, water years 1993–2017*. No. 2021-5079. US Geological Survey, 2021