Water Quality Prediction using Statistical, ML and Hybrid models

Shriya B	185001149
Vikram V	185001194
Vyshali S	185001202

BE CSE, Semester 7

Dr. D.Venkata Vara Prasad
Supervisor

Project Review: 0 (26 October 2021)
Department of Computer Science and Engineering
SSN College of Engineering

1 Introduction

Water is an essential elixir for several living organisms to function and survive. India contributes to 4 percent of the world's total freshwater resources. Out of the 4 percent Tamil Nadu contributes only 2.5 percentage. The ratio of population versus the freshwater availability is concerning. With a growing population, availability of good quality water is of grave importance. Water gets contaminated through sources such as industrial wastes, oil spills, marine dumping, pesticides, chemical fertilizers, waste water, sewage and etc. Water quality of resources has been analysed by several researchers before. Several models using Machine Learning, Deep Learning, Auto-ML(Machine Learning) and Auto-DL(Deep Learning) have been proposed before for the same. The objective of this research is to overcome the shortcomings of the previous models. This paper analyses water quality from the outcomes of statistical models and Machine learning models. Further, a hybrid model combining Statistical and Machine Learning models is proposed.

2 Problem statement

Water quality analysis has always been a major area of research. The proposed system will use a combination of statistical and Machine learning models. Real-world data is generally incomplete, inconsistent and noisy. The statistical model pre-processes the data set in order to resolve the shortcomings of real world data. Then, the Machine Learning model predicts the water quality of the water sample. Using these features, the ML model predicts the class feature (water quality). The results of the conventional statistical, machine learning models and the hybrid model is compared. The outcome of the hybrid model is compared with machine learning and statistics based system in order to analyse the performance of the hybrid model.

3 Justification for the problem formulation

Usage of statistics for pre processing on datasets used to find water quality is a new unexplored area. The proposed model will dig deep into this area and find the best method for pre processing. It will then predict the water quality of the sample using machine learning algorithm. This research also develops a statistical model and a Machine Learning model(Random Forest). Statistical techniques such as Linear Regression, Classification, and Unsupervised Learning Algorithms[PCA(Principal modelling techniques), Hierarchial clustering] are studied and the best Statistical model is taken for comparison with other models. Since there is not much research releated to this area, the best statistical model as well as the best statistic technique for pre processing can only be found after several trials. Research has been done to predict water quality analysis using standalone statistical and machine learning models. The objective of this work is to propose a new model to combine both statistical and machine learning models to predict water quality analysis.

4 Literature survey and Feasibility study

[1] using the data set from Perak River Basin, Malaysia proposed multiple neural networks which proved to be more accurate in classifying data set into its WQC(water quality class) with an accuracy of 85 percent. FANN(Feed Forward Artificial Neural Network) provided non-robust nature of prediction while MNN(multiple neural network) introduced robustness into the picture. The values of the coefficient of determination and MSE prove that polynomial regression and gradient boosting helped increase accuracy. The metrics only showed slight changes between the models of FANN and MNN.

[2] used machine learning and deep learning techniques such as SVM(Support Vector Machine), LDA(Linear Discriminant Analysis), LSTM(Long Short-Term Memory), DNN(Deep Neural Network), ANN(Artificial Neural Network), RNN(Recurrent Neural Network). They predicted water quality by proposing a new model least squares

support vector machine(LS-SVM). The dataset was from Liuxi river, Guangzhou from which eight features were used. This paper has shown that SVM does not work well on time series dataset. But after scaling the dataset, the prediction of SVM classifier has improved. It shows that there is a need to find better models for imbalanced datasets since all real world data are imbalanced.

[3] developed two machine learning models - the long short-term memory(LSTM) and Convolutional Neural Network(CNN) along with two deep learning models - support vector regression(SVR) and decision tree(DT). A coupled CNN-LSTM model was also developed. According to the statistical metrics LSTM outperformed CNN for DO prediction and DL models yielded similar results for Cha-a prediction. The hybrid model - integration of LSTM and CNN models outperformed both ML and DL standalone models in DO and Cha-a prediction. The dataset was from Small Prespa Lake, Greece. Statistical metrics such as correlation coefficient, root mean square error(RMSE), mean absolute error(MSE) and others were used to assess the performance of the models. The hybrid model captured both low and high levels of water quality variables for DO concentrations.

[4] used different machine learning models such as decision tree, random forest, logistic regression, support vector machine and navie bayesian network for binary and multi class classification. Random forest was the beat performing model among them with a accuracy of 96 percent. The dataset was acquired from Korattur Lake, Tamil Nadu. The machine learning models handle less data than what a hybrid model or a deep learning model would use. Increase in the dataset will lead to better results. So, there is a need for deep learning or hybrid models.

[5] used data obtained from National Water Information System(NIWS) to develop a Artificial Neural Network and Nonlinear Autoregressive time-series analysis with SCG(scaled conjugate gradient) for training algorithm. The performance metrics used for determining accuracy of model are Mean-Squared Error(MSE), Root Mean-Squared Error(RMSE) and Regression Analysis. The regression value for specific conductance was found to be 0.99. It states that the water quality analysis must be more user centric so that a solution for helping the environment could be found.

[6] compares deep learning techniques using unsupervised learning and supervised learning such as ANN techniques for predictive analysis of water quality on the data set from Chaskaman reservoir, Nasik. The performance metrics such as Mean Squared Error and Mean Absolute Error were used for unsupervised learning. The research has shown that using unsupervised learning techniques, prediction of data with variation has a good accuracy rate than supervised techniques.

[7] and many groups of researchers used multivariate statistical techniques such as cluster analysis(CA), principal component analysis(PCA), factor analysis(FA) and discriminant analysis(DA). [8] employed PCA(principal component analysis), CA(component analysis) and Pearson's correlation on dataset acquired from Morocco river. PCA(principal component analysis) and CA(component analysis) found sources for contamination of water for different seasons. This paper emphasises the importance of multivariate statistical assessment of large datasets to get better information.

[9]compared the pre-processing methods decimal scaling, min-max normalization and z-score normalization statistical techniques. It was found that z-score technique worked the best as the size of the dataset increases. The research also compared machine learning models such as HMCM(Hidden Markov chain model), Box-Cox Transformation and Linear Transformation out of which HMCM(Hidden Markov chain model) has the highest accuracy rate, prediction ability and utility. It shows that HMCM is the best statistical machine learning algorithm to use for big data analysis. A high dimensional data analysis technique needs to be integrated along with this model to find why this model performed the best.

[10] designed a pre-processing model based on mathematical statistics methods using quartile detection and Z-score method. The original data obtained from electricity consumption in Shanghai is abnormal and has many errors. The statistical methods extract required data, expel abnormal data and structure the dataset thus increasing the quality of the dataset. The model can be developed more by summarizing more features so that data classification in pre processing stage is easier and precise.

The data set was acquired from Korattur Lake which is to the north of Chennai-Arakkonam railway. The samples for the data set were collected from the lake every month for a span of 10 years. The Korattur lake data set consists of several features such as pH, Turbidity, TDS, Phosphate, Nitrate, Iron, COD, Chlorine, and Sodium. There is a binary classified data set as well as a multi-class data set. The dataset, hardware and software tools used for the execution of the research are available . So the work can be completed in the stipulated time.

4.1 Statistical Models

Multivariate statistical techniques examine relationships among multiple variables at the same time.

Correlation Analysis: It calculates summary statistics for each variable as well as correlation and covariance between variables.

Principle Component Analysis: It derives linear combinations of multiple quantitative variables that explain the largest percentage of the variation amongst those variables. These types of analyses are used to reduce the dimensionality of the problem in order to better understand the underlying factors affecting those variables.

Hierarchical Cluster Analysis: It begins with separate observations and groups them together based upon the distance between them in a multivariate space

Discriminant Analysis: The Discriminant Analysis procedure is designed to help distinguish between two or more groups of data based on a set of p observed quantitative variables.

4.2 Machine Learning Model - Random Forest

[11],[12]Random Forest is a Supervised Machine Learning Algorithm for solving problems on classification and regression. The majority vote after building decision trees on different samples is taken for classification. It can handle categorical values and performs better for classification rather than regression. It is a example for bagging type of ensemble technique. It creates a different training subset from sample training data with replacement & the final output is based on majority voting.

The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model.

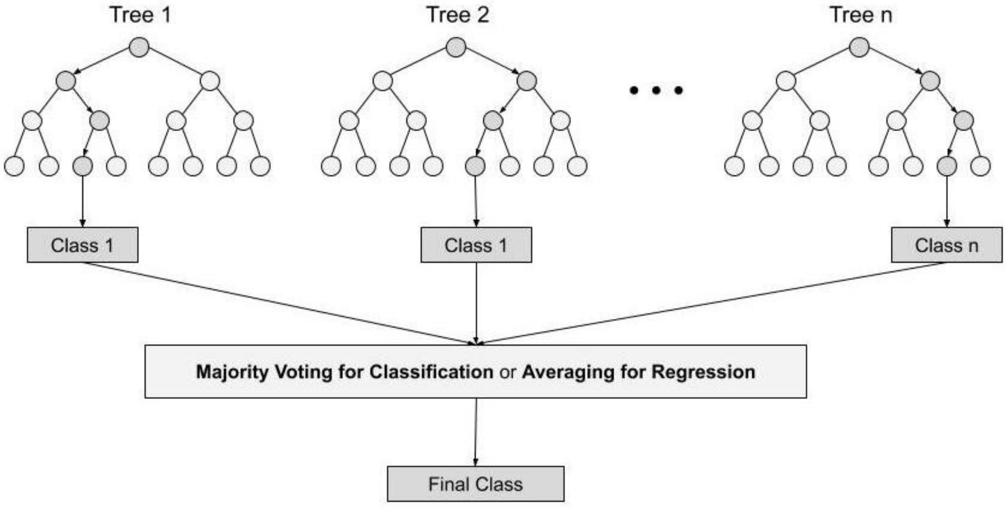
Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

Figure 1: Random Forest Working

5 Proposed system

Initially, the results obtained from Water quality analysis using a machine learning model (Random Forest) and a statistical model (such as PCA, CA, DA etc) are recorded. Later,the proposed model combining both statistical and machine learning models is deviced. This system contains a data pre-processing unit that processes the data set based on several parameters using a statistical model. Some of the statistical models include Z-score normalization, min-max normalization, decimal scaling and quartile detection method. Later this system uses a machine learning model (eg. Random Forest) that would predict the water quality. Finally, the results of all the three models are compared to check their performance measure.

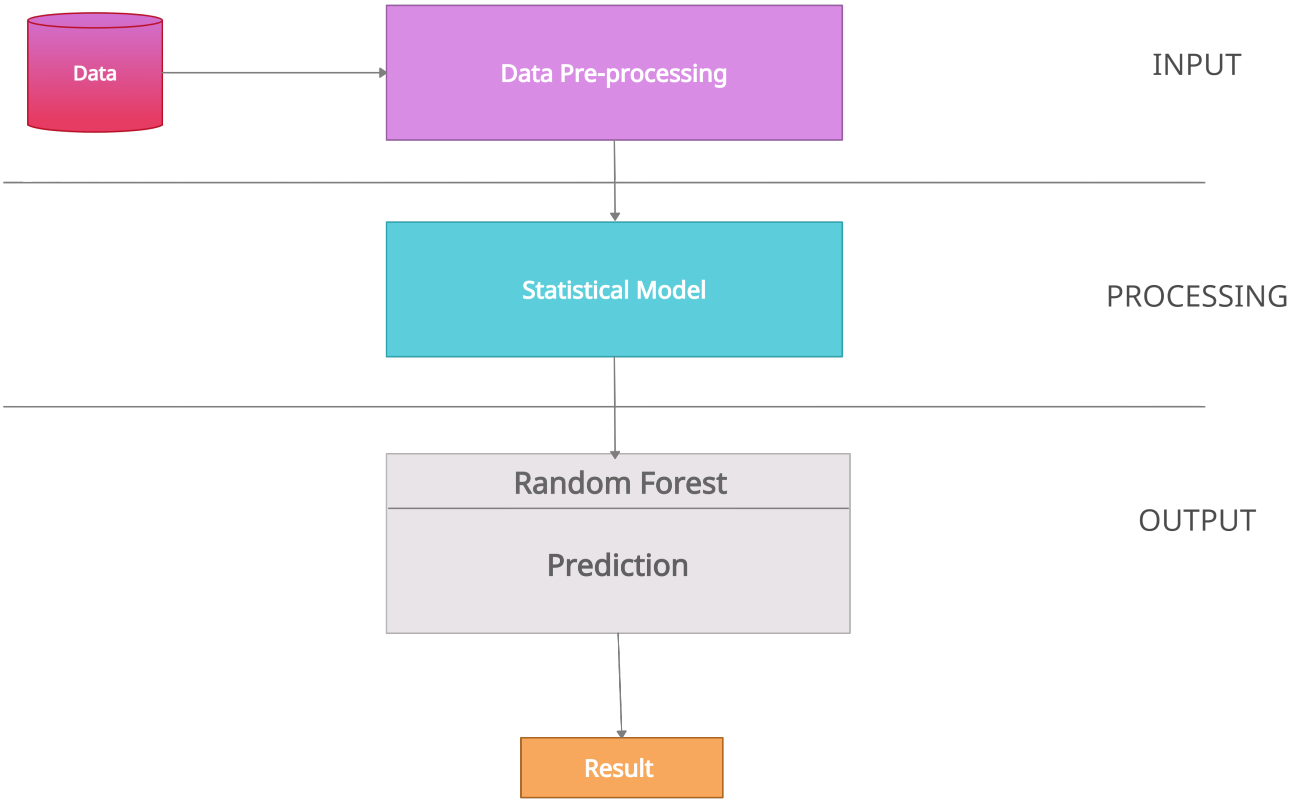

Figure 2: Hybrid model Architecture

6 Modules Split-up

This project comprise of four major modules:

1. ML module
2. Statistical module

3. Hybrid module

4. Result analysis

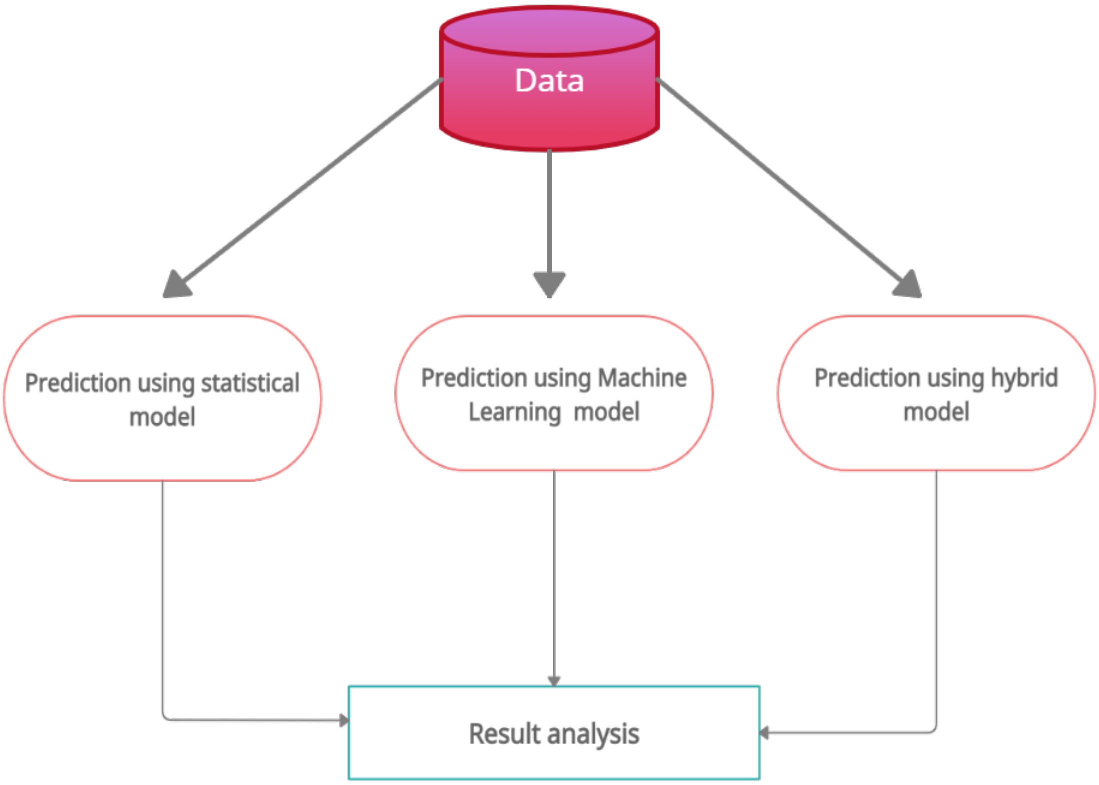
Data
Prediction using statistical model
Prediction using Machine Learning model
Prediction using hybrid model
Result analysis

Figure 3: Module Split up

References

[1] Ahmad, Z.; Rahim, N. A.; Bahadori, Alireza; Zhang, Jie (2017). Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. International Journal of River Basin Management, 15(1), 79–87. doi:10.1080/15715124.2016.1256297 .

[2] Fitore Muharemi, Doina Logofătu Florin Leon (2019): Machine learning approaches for anomaly detection of water quality on a real-world data set, Journal of Information and Telecommunication, https://doi.org/10.1080/24751839.2019.1565653.Muharemi, Fitore; Logofătu,

Doina; Leon, Florin (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication, (), 1–14. doi:10.1080/24751839.2019.1565653 .

[3] Barzegar, R., Aalami, M.T., Adamowski, J., 2020. Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model.Stoch. Environ. Res.Risk Asses.34,415-433. https://doi.org/10.1007/s0047-020-01776-2

[4] D. Venkata Vara Prasad , Lokeswari Y. Venkataramanaa , P. Senthil Kumarb, G. Prasannamedhab, K. Soumyaa, A.J. Poornemaa. 2021. Prediction on water quality of a lake in Chennai, India using machine learning algorithms. 218, 44-51. doi: 10.5004/dwt.2021.26970.

[5] Khan, Y., See, C.S., 2016. Predicting and analyzing water quality using Machine Learning: A comprehensive model. In: 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT) pp(1-6). DOI:10.1109/LISAT.2016.7494106

[6] Solanki, Archana, Agrawal, Himanshu, Khare, Kanchan, 2015.Predictive Analysis of Water Quality Parameters using Deep Learning. International Journal of Computer Applications (0975 – 8887) Volume 125 – No.9, 29-34.

[7] Muangthong, Somphinith; Shrestha, Sangam (2015). Assessment of surface water quality using multivariate statistical techniques: case study of the Nampong River and Songkhram River, Thailand. Environmental Monitoring and Assessment, Environ Monit Assess (2015) 187:548. doi:10.1007/s10661-015-4774-1

[8] Ahmed Barakata, Mohamed El Baghdadi, Jamila Rais, Brahim Aghezzaf, Mohamed Slassi, 2016. Assessment of spatial and seasonal water quality variation of Oum Er Rbia River (Morocco) using multivariate statistical techniques. International Soil and Water Conservation Research Volume 4, Issue 4, December 2016, Pages 284-292. https://doi.org/10.1016/j.iswcr.2016.11.002

[9] A. Rahman, Statistics-Based Data Preprocessing Methods and Machine Learning Algorithms for Big Data Analysis, International Journal of Artificial Intelligence, vol. 17, no. 2, pp. 44-65, 2019.

[10] M. Chen, Z. Huang, Q. Wu, W. Xu and B. Xiong, ”Pre-processing and audit of power consumption data based on composite mathematical statistics model,” 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), 2018, pp. 1-4, doi: 10.1109/EI2.2018.8582623.

[11] https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[12] https://builtin.com/data-science/random-forest-algorithmprocon