Water Quality Prediction using Statistical, Ensemble and Hybrid models

Vyshali S	185001202
Vikram V	185001194
Shriya B	185001149

BE CSE, Semester 8

Dr. D.Venkata Vara Prasad
Supervisor

Project Review: 1 (31 March 2022)
Department of Computer Science and Engineering
SSN College of Engineering

1 Title

Water Quality Prediction using Statistical, Ensemble and Hybrid methods.

2 Abstract

With a growing population, availability of good quality water is of grave importance. Water gets contaminated through several sources. Thus, the quality of water must be maintained so as to not risk human life.

The objective of this research is to analyse the data and predict the water quality of the resources by building a model with better prediction ability. The proposed Hybrid system will use a combination of statistical and Ensemble learning models. The statistical model pre-processes the data set in order to resolve the shortcomings of real world data. Statistical techniques such as Linear Regression, Classification, and Unsupervised Learning Algorithms [PCA(Principal modelling techniques), Hierarchical clustering] are studied and the best Statistical model is taken after comparison with other models. Then, the Ensemble Learning model predicts the quality of the water sample.

Ensemble methods create multiple models and combine them to produce better results.They usually produce solutions that are higher in accuracy than a single model. Bagging, Boosting and Stacking are the methods used in this research. These methods

are implemented using decision tree classifier, random forest, XGboost, K neighbours and logistic regression as base models. Bagging is generally used to reduce variance in a data set that contains noise. Boosting is a technique that creates a strong classifier from several weak classifiers. Stacking is a method that combines predictions of multiple models to create an optimal model. All three methods are fed with both binary and multi class data.

Finally, Bagging, Boosting and Stacking models are compared and the best model is selected to use for further stages. The statistical and ensemble learning models are combined to form a hybrid model. The outcome of the hybrid model is compared with ensemble learning and statistics based systems in order to analyse the performance of the hybrid model.

3 Architectural Design for Proposed system

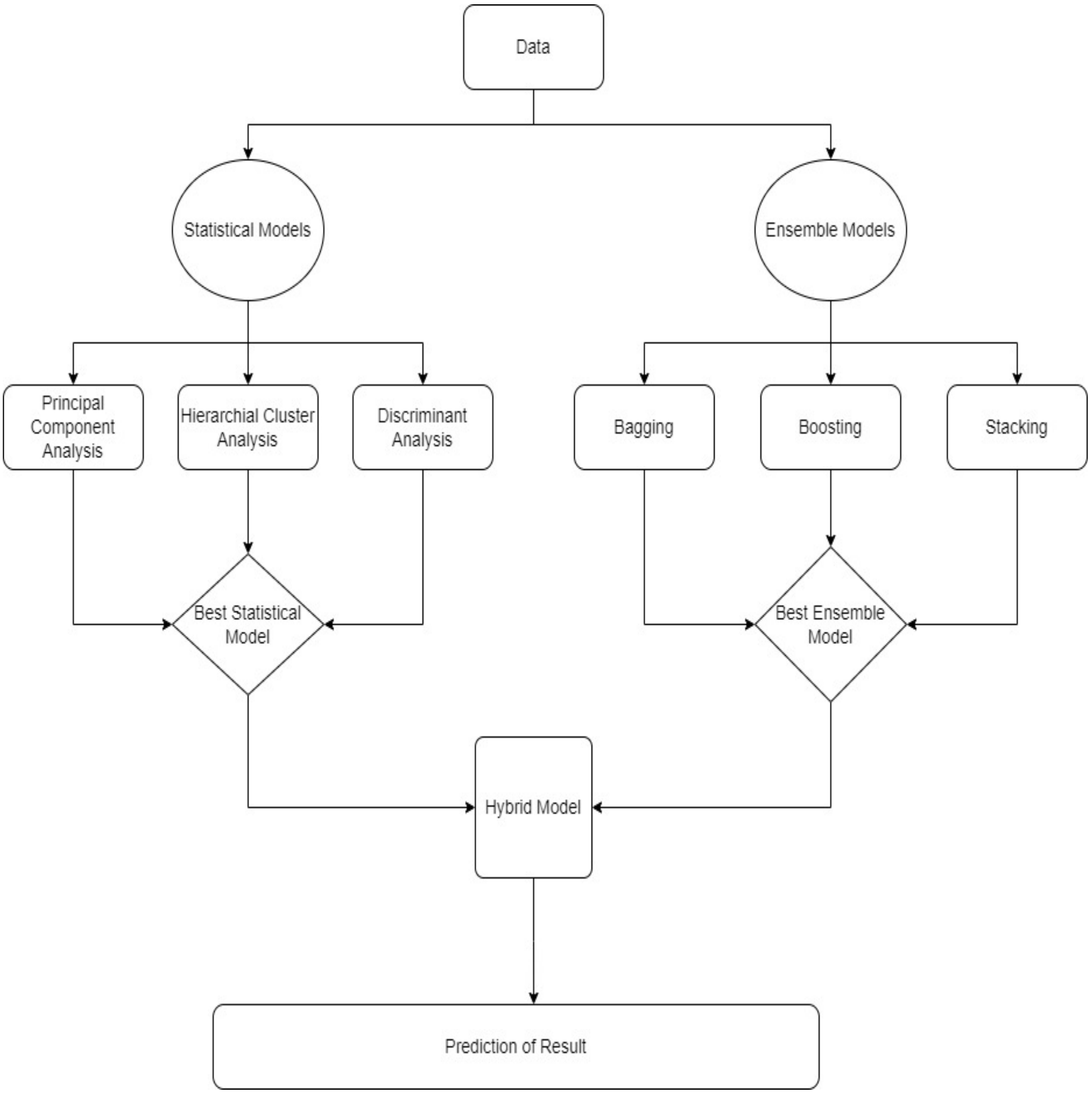


Figure 1: Proposed system Architecture

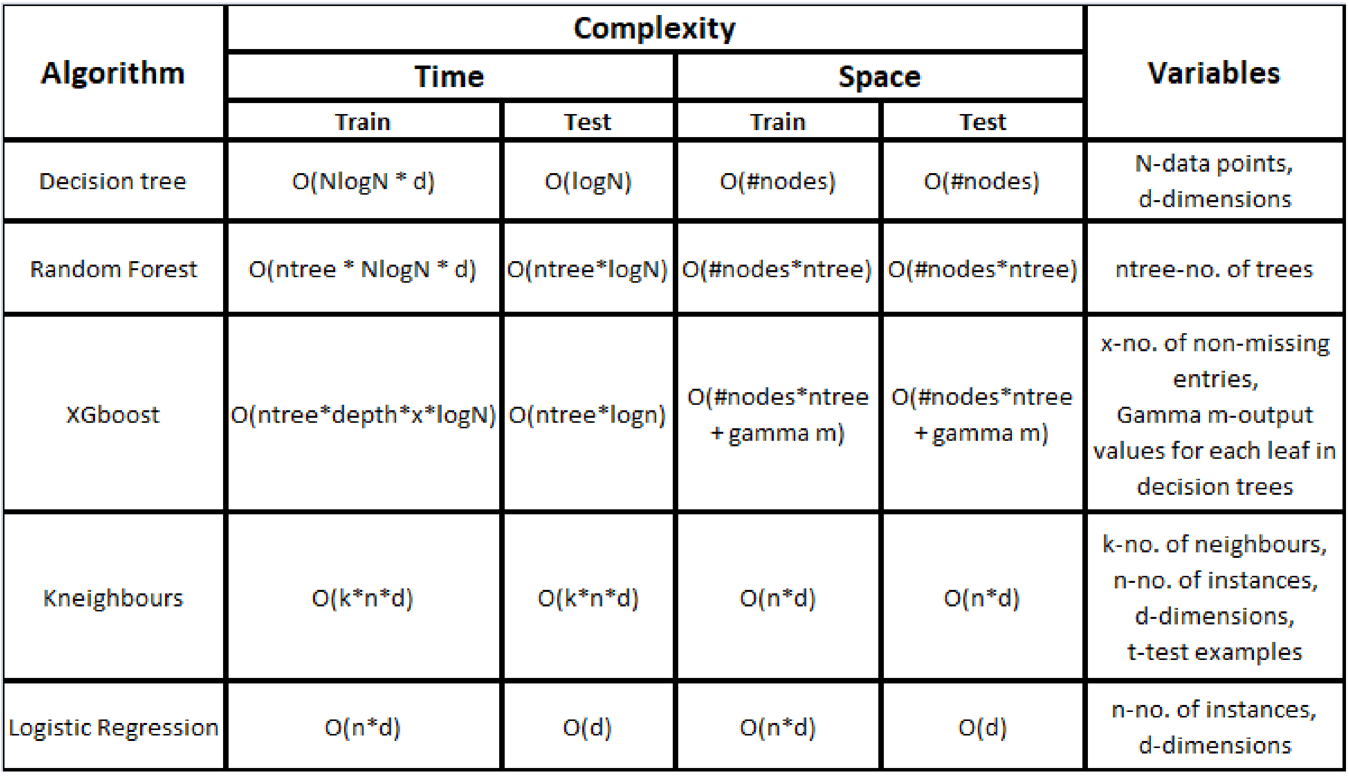
The data-sets - both binary class and multi class are fed into the ensemble models. The ensemble models taken for research are bagging, boosting and stacking. The base models such as decision trees, random forest, XGBoost, K neighbours and Logistic regression.

- **Decision trees** are found to **work best for Bagging Ensemble Model** with better accuracy than others.
- In **boosting**, both Ada-boost and XGBoost were modelled and **Ada-boost** was

found to be the **better model** among the two.

- In **stacking**, decision tree classifier, random forest, XGboost, K neighbours are used as base models along with linear regression as the final estimator.
- For **Binary class data**, comparing accuracy, precision, recall, F1 score, cross validation scores and training time, it was found that **Ada-boost works best** among all three models.
- For **Multi class data**, comparing the above mentioned scores as well as training time, **Bagging works best** among the 3 models.

4 Algorithms / Techniques used with complexity



Algorithm	Complexity				Variables
	Time		Space		
	Train	Test	Train	Test	
Decision tree	O(NlogN * d)	O(logN)	O(#nodes)	O(#nodes)	N-data points, d-dimensions
Random Forest	O(ntree * NlogN * d)	O(ntree*logN)	O(#nodes*ntree)	O(#nodes*ntree)	ntree-no. of trees
XGboost	O(ntree*depth*x*logN)	O(ntree*logn)	O(#nodes*ntree + gamma m)	O(#nodes*ntree + gamma m)	x-no. of non-missing entries, Gamma m-output values for each leaf in decision trees
Kneighbours	O(k*n*d)	O(k*n*d)	O(n*d)	O(n*d)	k-no. of neighbours, n-no. of instances, d-dimensions, t-test examples
Logistic Regression	O(n*d)	O(d)	O(n*d)	O(d)	n-no. of instances, d-dimensions

Figure 2: Base models

4.1 Base models

4.1.1 Decision tree

A decision tree is a decision support tool that uses a tree-like model of **decisions and their possible consequences**, including chance event outcomes, resource costs, and utility. Their internal nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the outcome.

4.1.2 Random Forest

Random forest is an ensemble method that is made up of a large number of **small decision trees**, called **estimators**, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

4.1.3 XGBoost

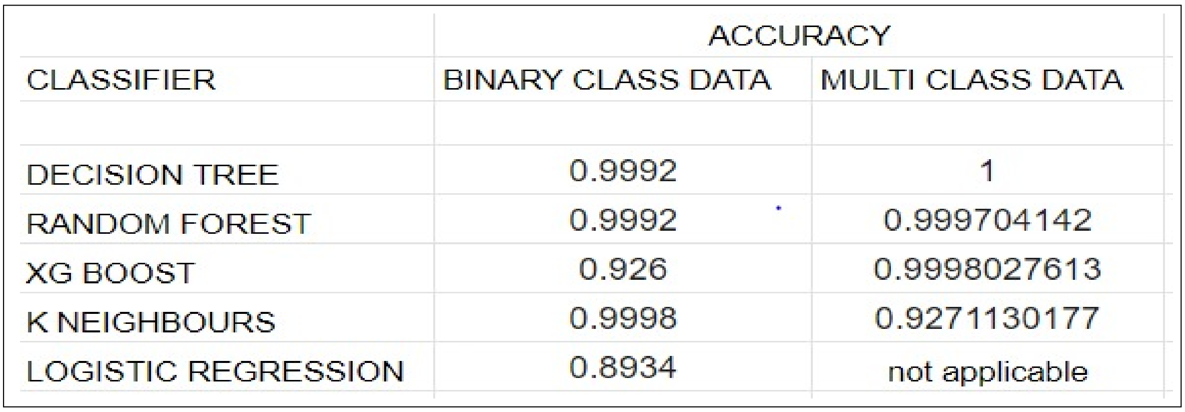
XGBoost, which stands for **Extreme Gradient Boosting**, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides **parallel tree boosting** and is designed to be highly efficient, flexible and portable.It implements Machine Learning algorithms under the Gradient Boosting framework.

4.1.4 K-neighbours

The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the **likelihood** that a data point will become a **member of one group** or another **based on** what group the data points **nearest** to it belong to.It is a **lazy learning** and **non-parametric algorithm**.

4.1.5 Logistic regression

Logistic regression is used to **predict a dependent categorical target variable**.The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.



	ACCURACY	
CLASSIFIER	BINARY CLASS DATA	MULTI CLASS DATA
DECISION TREE	0.9992	1
RANDOM FOREST	0.9992	0.999704142
XG BOOST	0.926	0.9998027613
K NEIGHBOURS	0.9998	0.9271130177
LOGISTIC REGRESSION	0.8934	not applicable

Figure 3: Base Models Accuracy

4.2 Ensemble Techniques

4.2.1 Bagging

In parallel methods we fit the different considered learners independently from each other and, so, it is possible to train them concurrently. The most famous such approach is “bagging” (standing for “bootstrap aggregating”) that aims at producing an ensemble model that is more robust than the individual models composing it.

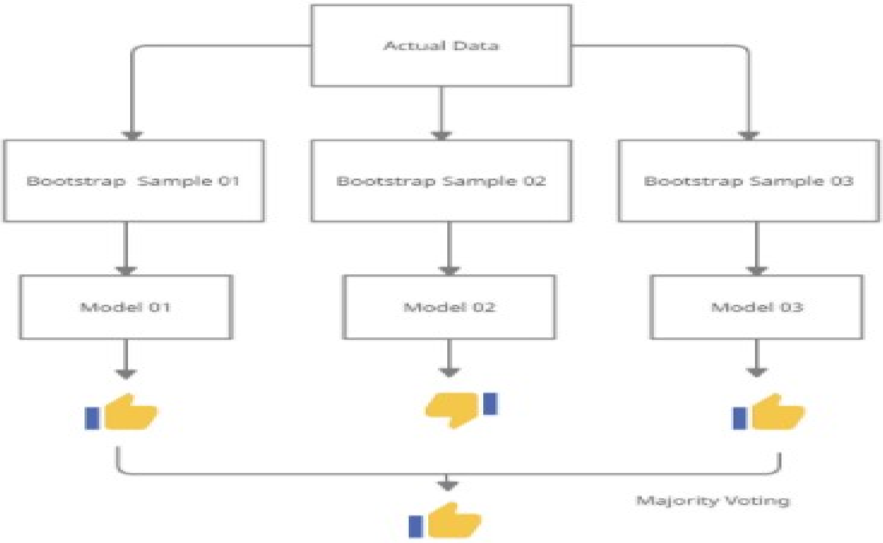


Figure 4: Bagging

4.2.2 Boosting

In sequential methods the different combined weak models are no longer fitted independently from each other. The idea is to fit models iteratively such that the training of models at a given step depends on the models fitted at the previous steps. “Boosting” is the most famous of these approaches and it produces an ensemble model that is in general less biased than the weak learners that compose it.

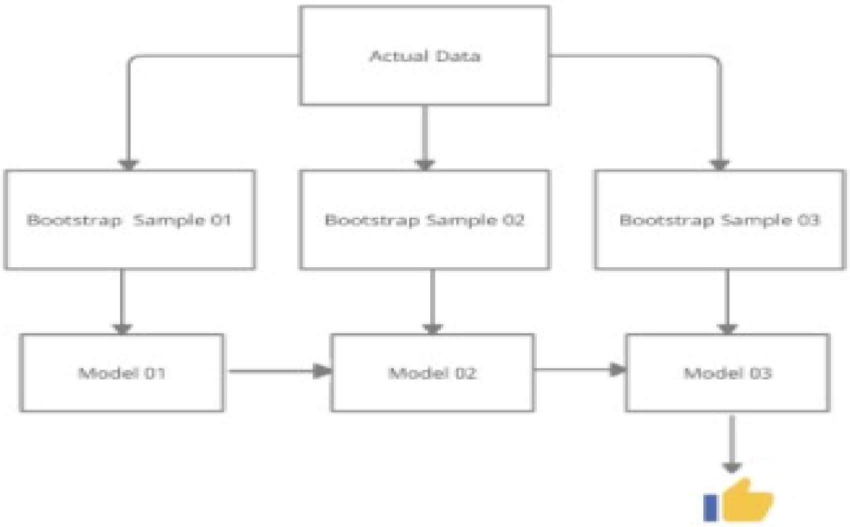


Figure 5: Boosting

4.2.3 Stacking

The idea of stacking is to learn several different weak learners and combine them by training a meta-model to output predictions based on the multiple predictions returned by these weak models. So, we need to define two things in order to build our stacking model: the L learners we want to fit and the meta-model that combines them.

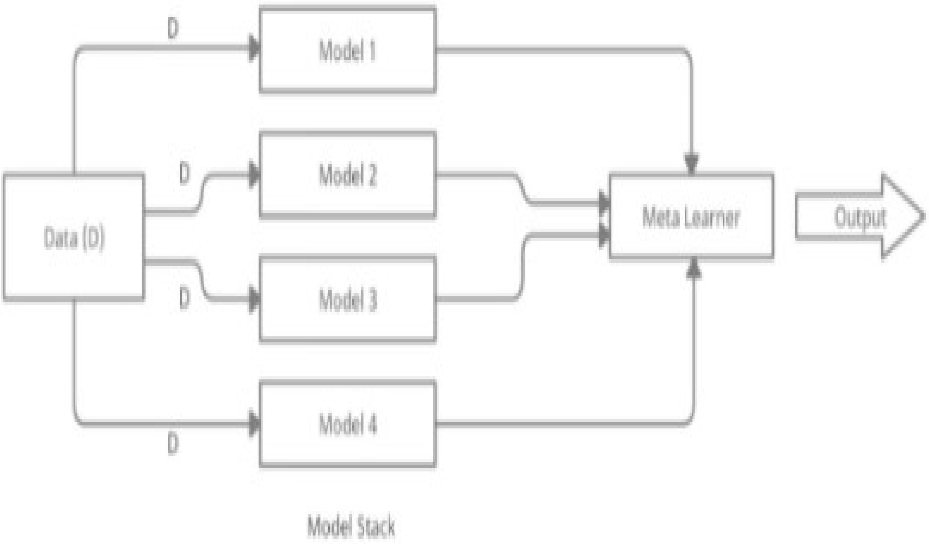


Figure 6: Stacking

5 Results

5.1 Exploratory Data Analysis

Classification of the records :

- Binary - Classes 0 & 1
- Multi - Classes 0, 1 & 2

Both the binary and multi class datasets contain no outliers, null values and noise.

5.1.1 Binary Class Data

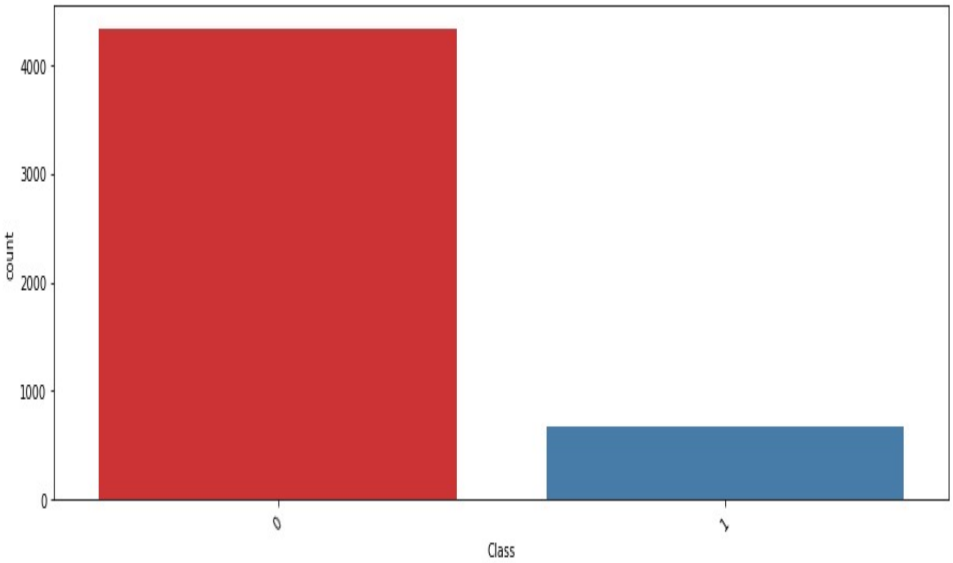


Figure 7: Class Division

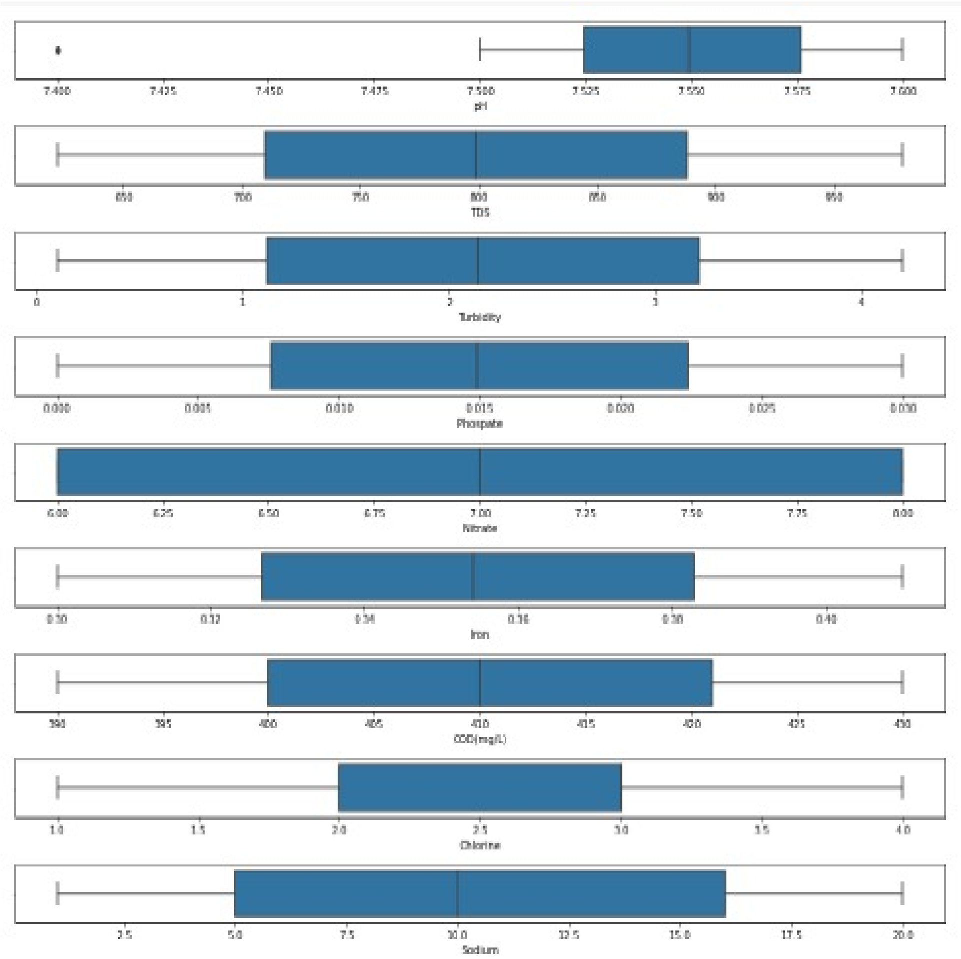


Figure 8: Boxplot of Binary class data

5.1.2 Multi Class Data

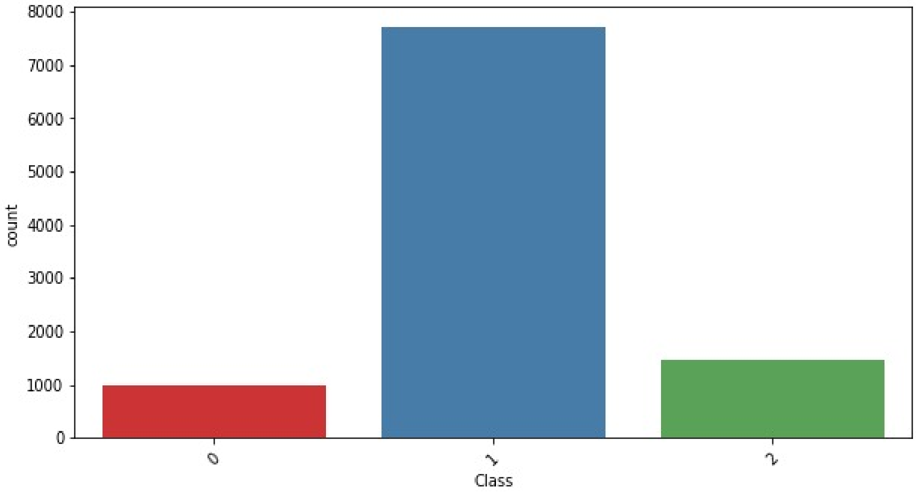


Figure 9: Class Division

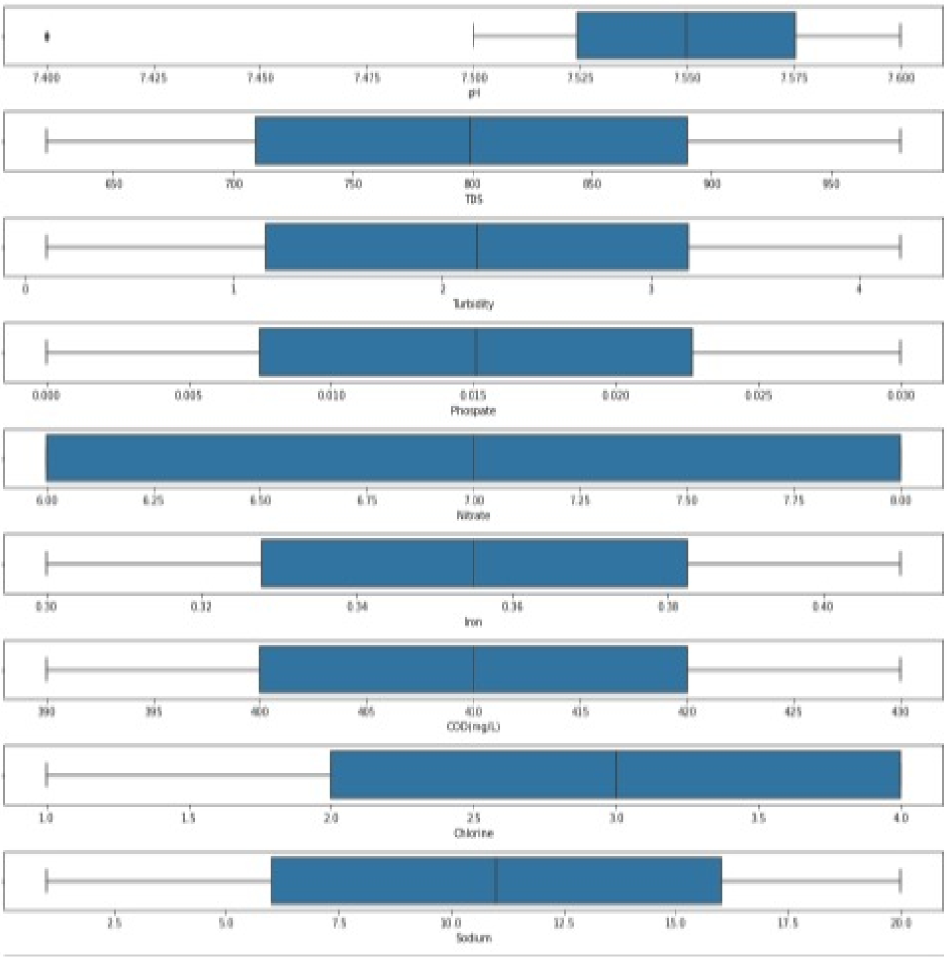


Figure 10: Boxplot of Multi class data

5.2 Model Training

- Three Ensemble models were trained - Bagging, Boosting and Stacking.

- The Bagging Classifier was tried with different base models such as decision tree classifier, random forest, XGBoost, K neighbours and logistic regression. Bagging Model showed the highest accuracy when Decision Tree Classifier was used as its base model.

- Adaboost Classifier is a boosting method where Decision tree classification is used as the base model.

- Stacking Classifier is an ensemble method which uses decision tree classifier, random forest, XGBoost, K neighbours as its estimators and uses Logistic Regression model as its final estimator.

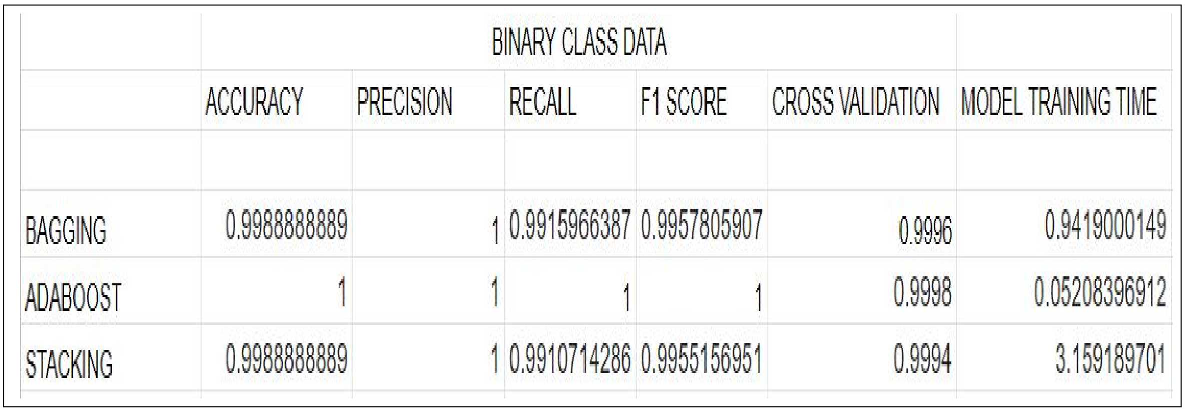
- The accuracy, precision, recall, F1 score and cross validation score have been calculated for each ensemble model.
- The time taken for training each model has also been estimated.
- The accuracy score of each base model has been calculated and compared with the ensemble models. It has been found that **ensemble models**, which are a combination of the base models, **have a higher accuracy than the base models** on their own.

.

5.3 Outcomes

5.3.1 Binary Class data

- For Bagging, an accuracy of 99.8% was achieved using Binary data.
- In the case of Adaboost, an accuracy of 100% was achieved and Stacking achieved an accuracy of 99.8%. While checking the precision, Bagging achieved 100% , Adaboost achieved the same precision as boosting and Stacking achieved 99.8%.
- The recall observed for Bagging was 99.2%. In case of Adaboost the recall observed was 100% and the Recall observed in stacking was 99.6%.
- The F1 Score achieved in Bagging was 99.6%, in boosting was 100%, and in stacking was 99.6%.
- Further on, the Cross validation observed was 99.9% for bagging, 99.9% for boosting and 99.94% for stacking.
- The model training time was 0.941, 0.052 and 3.159 seconds respectively for Bagging, Boosting and Stacking.



	BINARY CLASS DATA					
	ACCURACY	PRECISION	RECALL	F1 SCORE	CROSS VALIDATION	MODEL TRAINING TIME
BAGGING	0.9988888889	1	0.9915966387	0.9957805907	0.9996	0.9419000149
ADABOOST	1	1	1	1	0.9998	0.05208396912
STACKING	0.9988888889	1	0.9910714286	0.9955156951	0.9994	3.159189701

Figure 11: Binary class data outcome

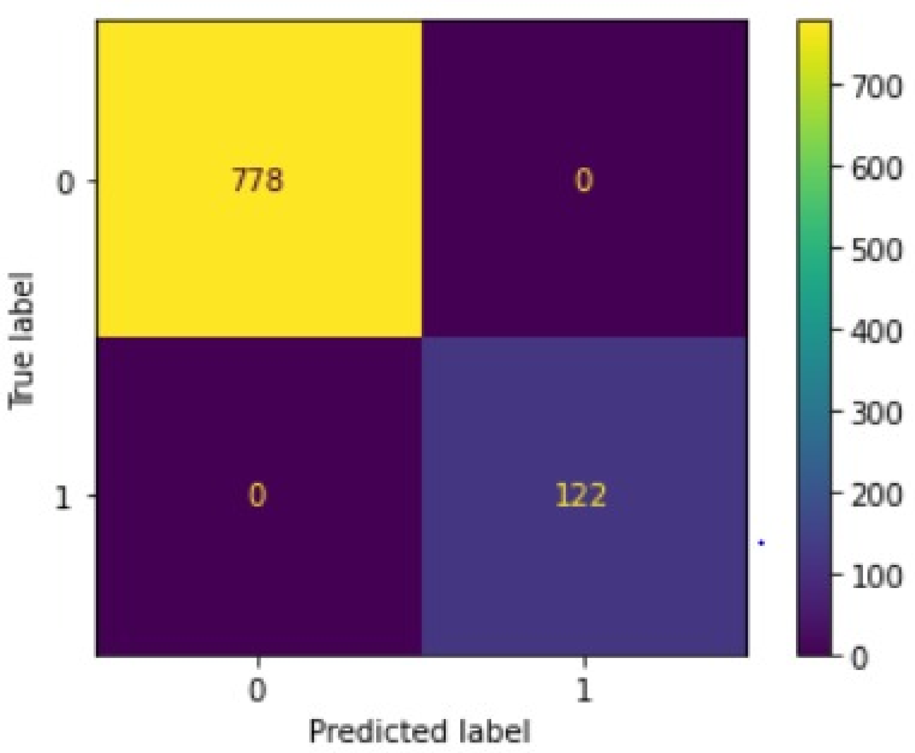


Figure 12: Confusion matrix for test Binary data - Bagging using Decision tree, Adaboost, Stacking

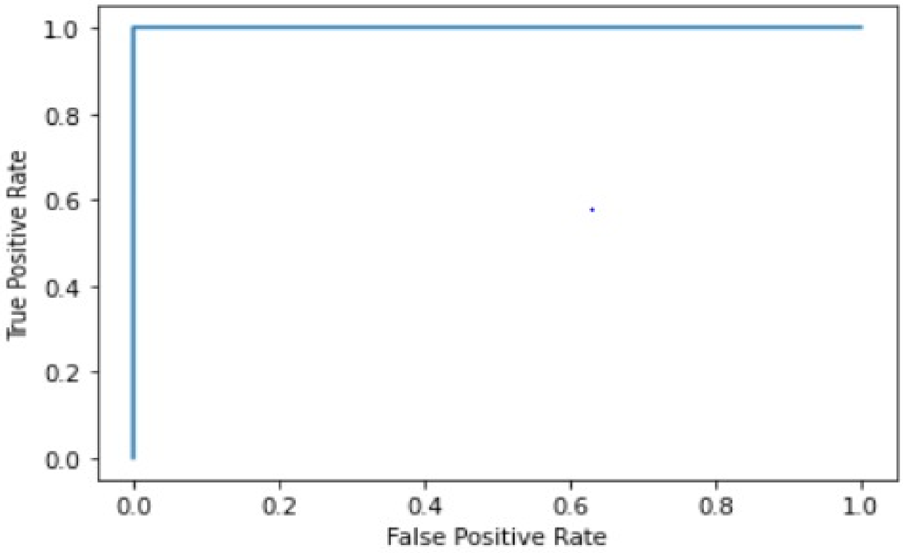


Figure 13: ROC curve for test Binary data - Bagging using Decision tree, Adaboost, Stacking

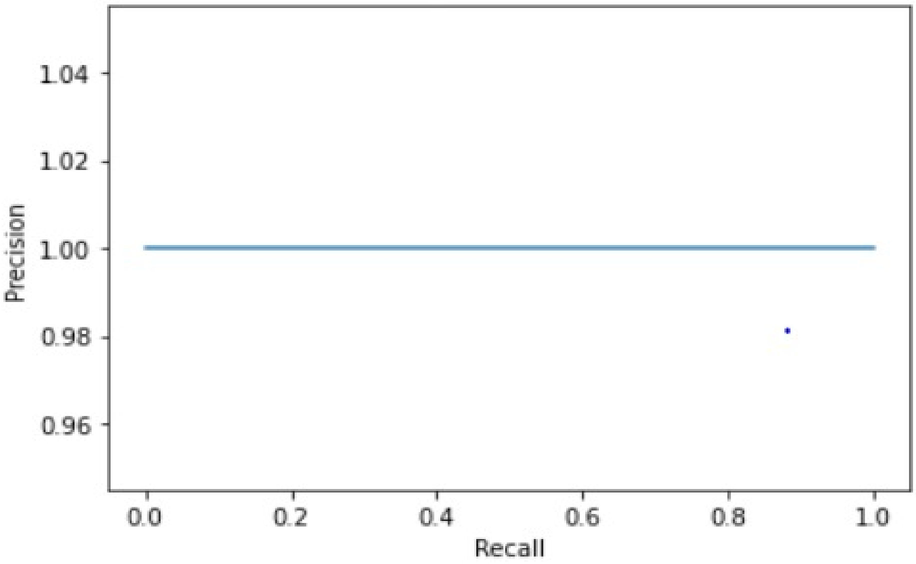


Figure 14: Precision-Recall display for test Binary data - Bagging using Decision tree, Adaboost, Stacking

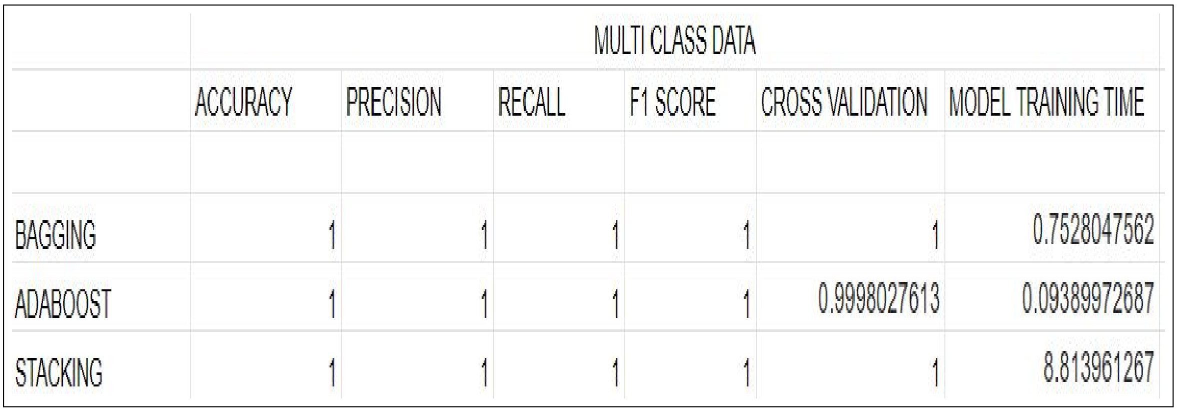
5.3.2 Multi Class data

- For Bagging, Boosting and Stacking, an accuracy of 100% was achieved using Multi class data.

- While checking the precision, Bagging, Boosting and Stacking achieved 100%

again.

- Similarly,The recall observed for Bagging, Boosting and Stacking was 100%.
- The F1 Score achieved in Bagging, Boosting and Stacking was 100% too.
- Cross validation observed was 100% for bagging and stacking, while the Cross validation observed was 99.98% for stacking.
- The model training time was 0.752, 0.093 and 8.813 seconds respectively for Bagging, Boosting and Stacking.



	MULTI CLASS DATA					
	ACCURACY	PRECISION	RECALL	F1 SCORE	CROSS VALIDATION	MODEL TRAINING TIME
BAGGING	1	1	1	1	1	0.7528047562
ADABOOST	1	1	1	1	0.9998027613	0.09389972687
STACKING	1	1	1	1	1	8.813961267

Figure 15: Multi class data outcome

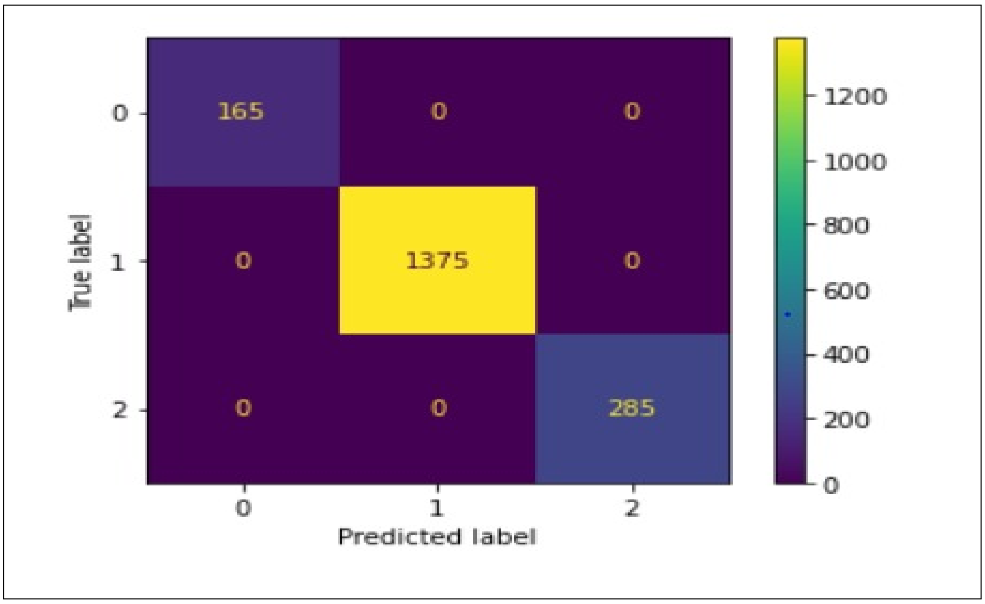


Figure 16: Confusion matrix for test Multi class data - Bagging using Decision tree, Ad- aboost, Stacking

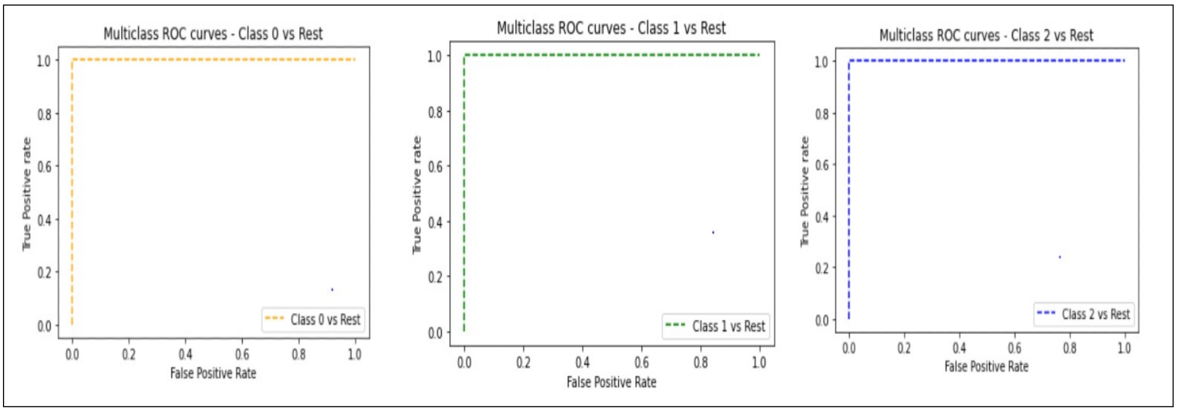


Figure 17: ROC for test Multi class data - Bagging using Decision tree, Ad- aboost, Stacking

**Model taining time is mentioned in seconds

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*, 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.* DOI : 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] Farid Hassanbaki Garabaghi, Semra Benzer, Recep Benzer, ”Performance Evaluation of Machine Learning Models with Ensemble Learning approach in Classication of Water Quality Indices Based on Different Subset of Features”, *Water resources management,Springer* November 3rd, 2021 DOI : https://doi.org/10.21203/rs.3.rs-876980/v1

[12] Victor Henrique Alves Ribeiro Nacre Capital, ”Monitoring of drinking-water quality by means of a multi-objective ensemble learning approach”, *The Genetic and Evolutionary Computation Conference Companion*, July 2019, DOI:10.1145/3319619.3326745

[13] https://medium.com/analytics-vidhya/computational-complexity-of-ml-algorithms-1bdc88af1c7a