Water Quality Prediction using Statistical, Ensemble and Hybrid models

Shriya B 185001149
Vikram V 185001194
Vyshali S 185001202
BE CSE, Semester 8

Dr. D.Venkata Vara Prasad
Supervisor

Project Review: 2 (22 April 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 such as industrial wastes, oil spills, marine dumping, etc. 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 machine learning models(using ensembling). 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. Statistical techniques such as Linear Regression, Classification, linear discriminant analysis (LDA), quadratic discriminant analysis 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.

In order to reduce dimensionality and noise from a real world dataset, statistical models are applied. Statistical models are mathematical techniques and statistical assumptions that generate sample data and make predictions. It usually is a collection of

probability distributions on a set of all possible outcomes of an experiment. The models used in this research are Principal Component Analysis, Hierarchical Clustering Analysis, Quadratic Discriminant Analysis and Linear Discriminant Analysis. Principal Component Analysis is a dimensionality reduction technique that reduces the dimension of a large data set while preserving the important information. Hierarchical Clustering Analysis technique clusters points that are more closely related to each other. Linear Discriminant Analysis used to find a linear combination of attributes that separates several classes of objects or events. Quadratic Discriminant Analysis is another version of LDA in which a separate covariance matrix is assumed for every class of outcomes. PCA, HCA QDA and LDA are compared and the best model is used for the data pre-processing.

Once the pre-processing of the data is done using the best performing statistical model, it is fed into the water quality prediction model. In order to determine the water quality, ensemble learning methods are used. 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 dataset 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 multiclass 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 System Architecture

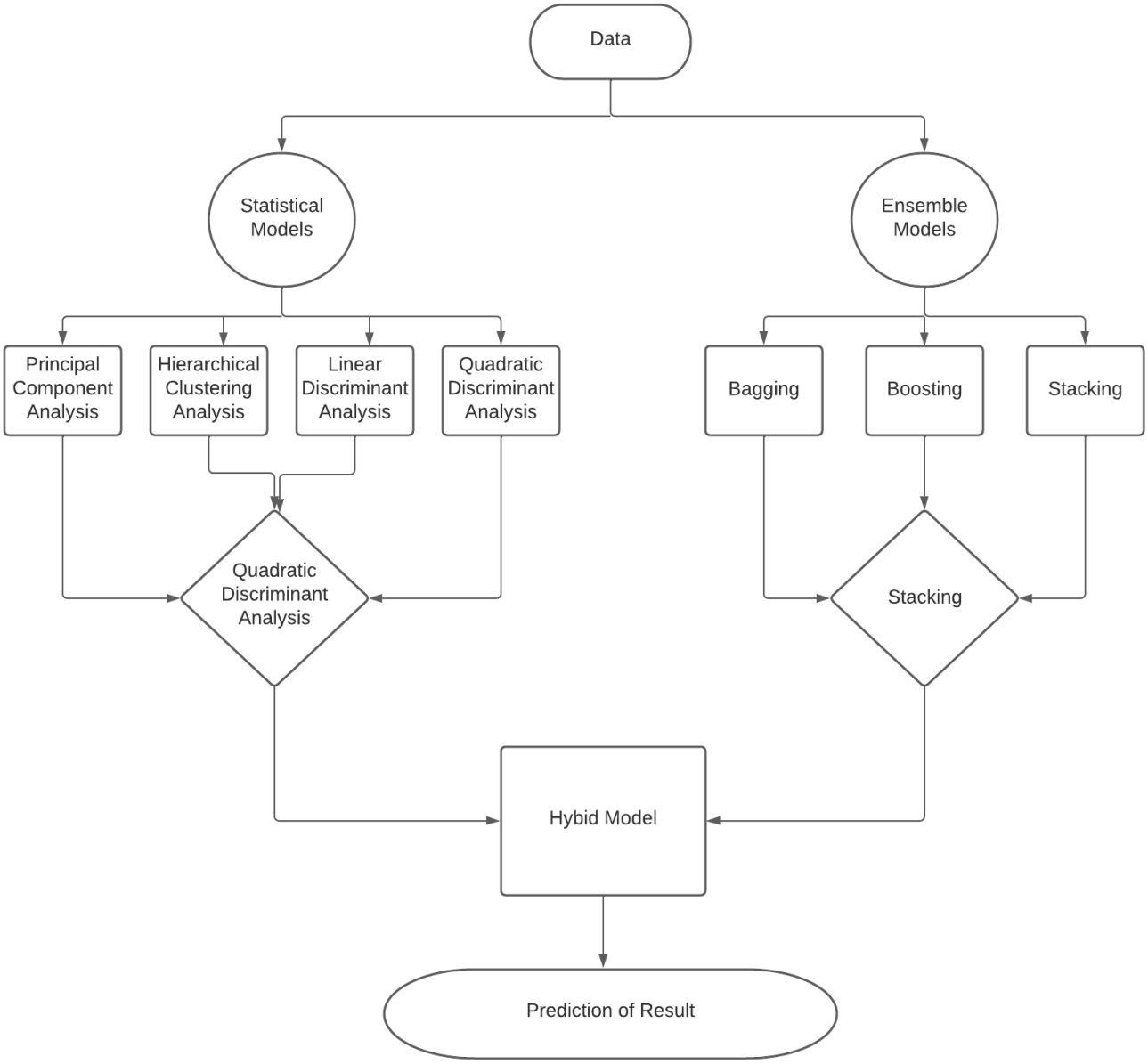

Figure 1: Proposed system Architecture

The datasets - both binary class and multi class are fed into both the statistical and ensemble models in parallel.
The Statistical Models implemented are Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis(QDA). Out of the four Statistical models, Quadratic Discriminant Analysis(QDA) was found to be the best performing model.

The Ensemble Learning models implemented are Bagging, Boosting and Stacking. Bagging was implemented using the base models decision trees, random forest, XGBoost, K neighbours and Logistic regression. Out of these, the Bagging model using Decision Trees performed best. Boosting model was implemented using AdaBoost and XGBoost and out of the two Adaboost was chosen due to its speed and higher accuracy. Stacking was implemented using the base models such as decision trees, random forest, XGBoost, K neighbours and Linear regression as the final estimator. Out of the three ensemble models, Stacking was chosen as the best model due to its performance. In the end the best statistical and ensemble learning models, that is Quadratic Discriminant Analysis and Stacking will be combined to form a new hybrid model which will perform better than its parent models when predicting the quality of water.

4 Algorithms

Algorithm	Time Complexity	Space Complexity	Note
PCA	$O(p^2n+p^3)$	O(nd)	n : number of samples, p : dimensions of a sample, d : number of variables
HCA	O(n2)	O(n3)	
LDA	$O(mnt + t^3)$	O(mn + mt + nt)	m : the number of samples, n : the number of features, t = min(m, n)

Figure 2: Statistical Models

4.1 Statistical Techniques

4.1.1 Principal Component Analysis

PCA is a common model used for **dimensionality reduction**, that is,reducing the feature space by removing several features from a real world dataset that is noisy and unclean. By removing these features, the dataset is made much **easier to visualise,**

analyse and interpret.PCA also **determines correlations between the features**. It is commonly used in the areas of pattern recognition and signal processing. There are two classes that come under PCA, namely- Feature Elimination and Feature Extraction.

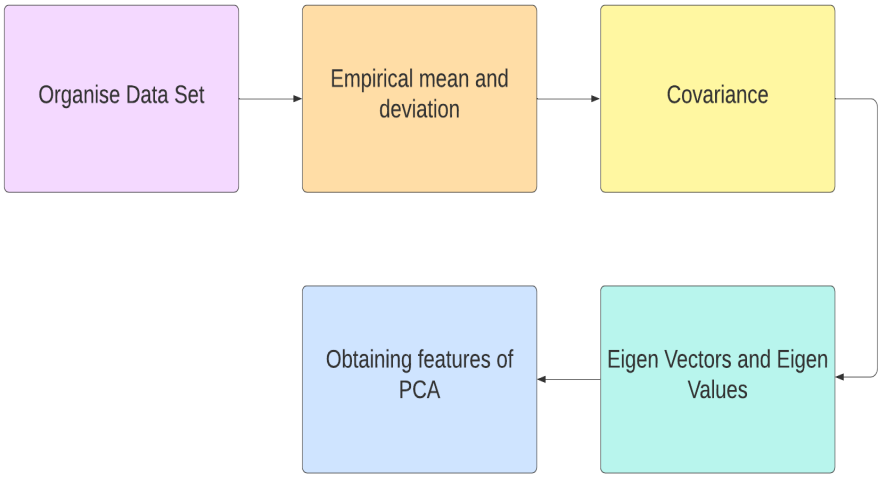

Figure 3: PCA

4.1.2 Hierarchical clustering Analysis

Hierarchical clustering analysis is a common model in which the objective is to **group several features/data points in such a way that they are close to one another**. The fundamental technique is to repeatedly calculate the distance between the features and further calculate the distances between the clusters once the features/ data points start forming clusters. The **outputs** are usually **represented** as a **dendrogram**. The two methods that fall under HCA are Divisive methods and Agglomerative methods.

S.no	Inputs
1	i1
2	i2
3	i3
4	i4
5	i5
6	i6

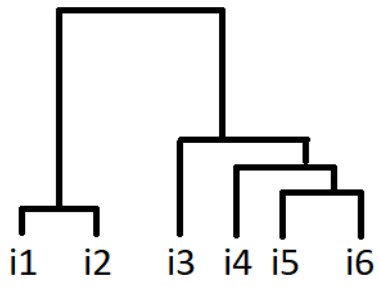

Figure 4: HCA

4.1.3 Linear Discriminant Analysis

Linear Discriminant analysis is also a commonly used **dimensionality reduction method** that is usually used in **supervised** classification problems. It is more precisely used to **model the differences between the groups/classes**. The higher dimension space is projected into the lower dimension space. Quadratic Discriminant analysis, flexible discriminant analysis and regularised discriminant analysis are the extensions to linear discriminant analysis. LDA is commonly used in the areas of medicine, face recognition, customer identification and so on.

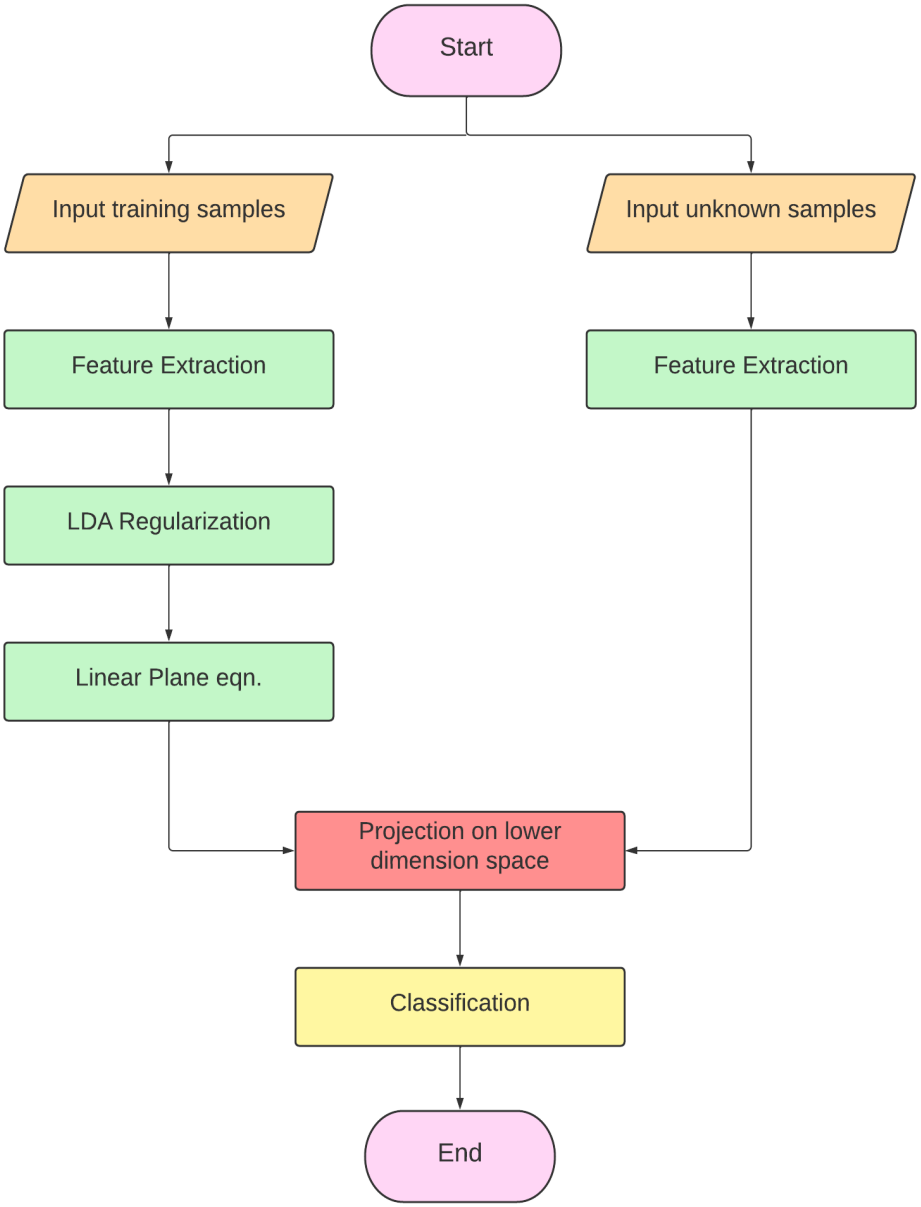

Figure 5: LDA

4.1.4 Quadratic Discriminant Analysis

QDA is quite related to linear discriminant analysis (LDA). It is assumed that the **measurements are normally distributed**. Unlike LDA, in QDA there is no assumption that the covariance of each of the classes is the same. QDA is a **generative model** and it assumes that every class follows a Gaussian distribution.One aspect in which the two differ is that LDA assumes the feature covariance matrices of both classes are identical, which leads to a linear decision boundary. However, QDA is less stringent.

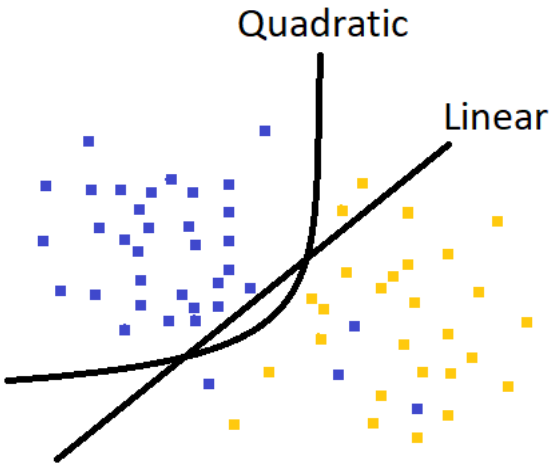

Figure 6: QDA vs LDA

4.2 Ensemble Techniques

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 7: Base models for 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.

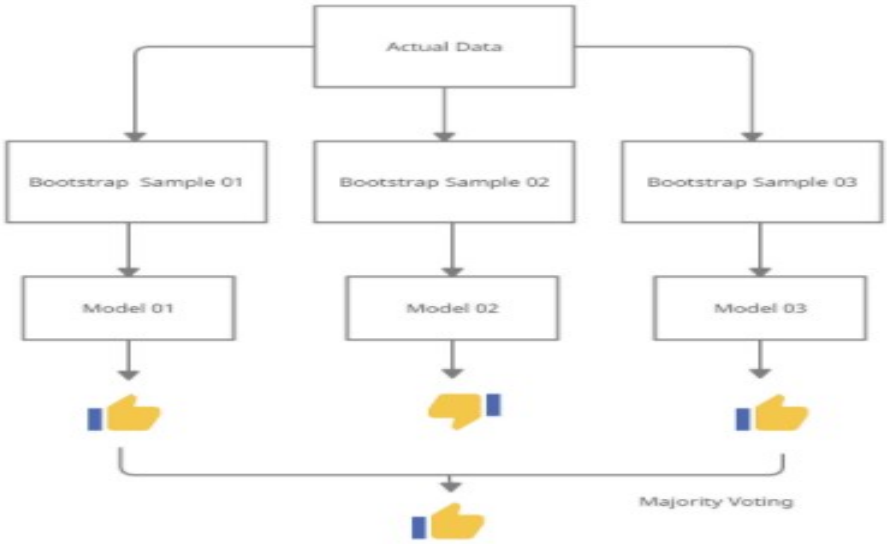

Figure 8: Bagging

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 9: 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.

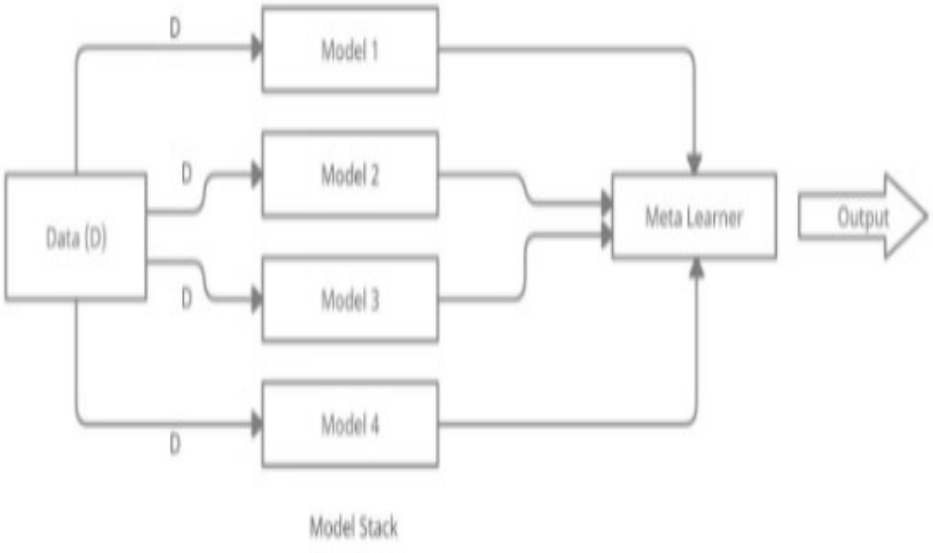

Figure 10: Stacking

5 Exploratory data analysis

5.1 Data-set Description

- This research compares the outcomes of the models using 3 different datasets, namely- Binary and Multiclass versions of Korattur Lake dataset and Kaggle dataset (Binary).
- The Binary Korattur lake dataset has two classes as the name suggests, where 0 indicates that the water is drinkable and 1 indicates that the water is non-drinkable. The dataset has 5001 rows and 10 columns.
- However, the multiclass version of the Korattur Lake dataset has 3 classes, where 0 indicates that the quality of water is excellent, 1 indicates that the water is good, and 2 indicates that the water quality is poor. The dataset has 10140 rows and 10 columns. There are 9 aspects based upon which the dataset is classified in both the Binary and Multiclass datasets. They are pH, TDS, Turbidity, Phosphate, Nitrate, Iron, COD(mg/L), Chlorine and Sodium.
- The Kaggle dataset with binary data has two classes,where 0 indicates that the water is unsafe and 1 indicates that the water is safe. There are 20 aspects based on which the dataset is classified. They are Aluminium, Ammonia, Arsenic, Barium, Cadmium, Cloramine, Chromium, Copper, Fluoride, Bacteria, Lead, Nitrates, Nitrites, Mercury, Perchlorate, Radium, Selenium, Silver and Uranium. There are 8000 rows and 21 columns in the dataset.

Dataset	Classes	Size (rows*colums)
Binary-Korattur	0(drinkable) & 1(non-drinkable)	5001*10
Multi-Korattur	0(excellent), 1(good) & 2(poor)	10140*10
Binary-Kaggle	0 (not safe), 1(safe)	8000*21

Figure 11: Classes, Size Description of all the data-sets

	pH	TDS	Turbidity	Phospate	Nitrate	Iron	COD(mg/L)	Chlorine	Sodium	Class
0	7.6	877	3.59	0.026136	8	0.378500	397	4	8	1
1	7.6	729	1.75	0.020622	6	0.333759	397	3	9	1
2	7.5	622	3.44	0.004071	7	0.382368	394	4	16	0
3	7.6	864	2.80	0.022071	7	0.313915	403	2	5	0
4	7.6	656	1.81	0.004031	7	0.333226	421	4	11	0

Figure 12: 1st 5 rows of Korattur Binary data-set

	pH	TDS	Turbidity	Phospate	Nitrate	Iron	COD(mg/L)	Chlorine	Sodium	Class
0	7.6	973	0.16	0.012967	8	0.328568	422	2	12	1
1	7.6	975	3.17	0.016066	7	0.332097	427	4	2	1
2	7.5	755	2.53	0.019433	7	0.338934	398	2	16	2
3	7.6	686	4.15	0.018559	7	0.303969	409	4	12	1
4	7.6	858	3.90	0.002456	8	0.383476	390	4	17	0

Figure 13: 1st 5 rows of Korattur Multi data-set

luminium	Ammonia	Arsenic	Barium	Cadmium	Cloramine	Chromium	Copper	Flouride	Bacteria	...	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	Class
1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.20	...	0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1
2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.90	0.65	...	0.100	2.01	1.93	0.003	32.26	3.21	0.08	0.27	0.05	1
1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	...	0.078	14.16	1.11	0.006	50.28	7.07	0.07	0.44	0.01	0
1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	...	0.016	1.41	1.29	0.004	9.12	1.72	0.02	0.45	0.05	1
0.92	24.33	0.03	0.20	0.006	2.67	0.69	0.57	0.61	0.13	...	0.117	6.74	1.11	0.003	16.90	2.41	0.02	0.06	0.02	1

Figure 14: 1st 5 rows of Kaggle Binary data-set

5.2 Distribution of records across classes

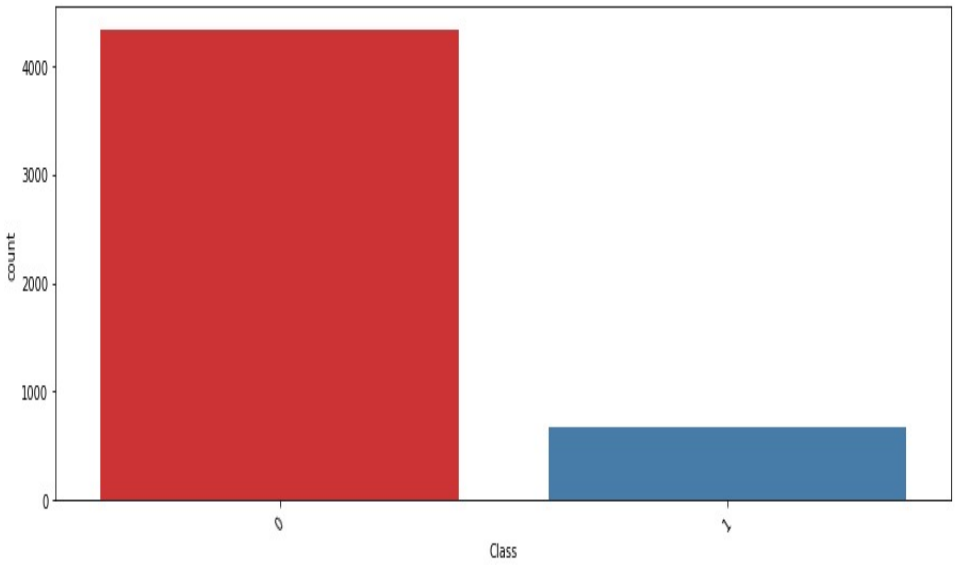

Figure 15: Binary Korattur data-set

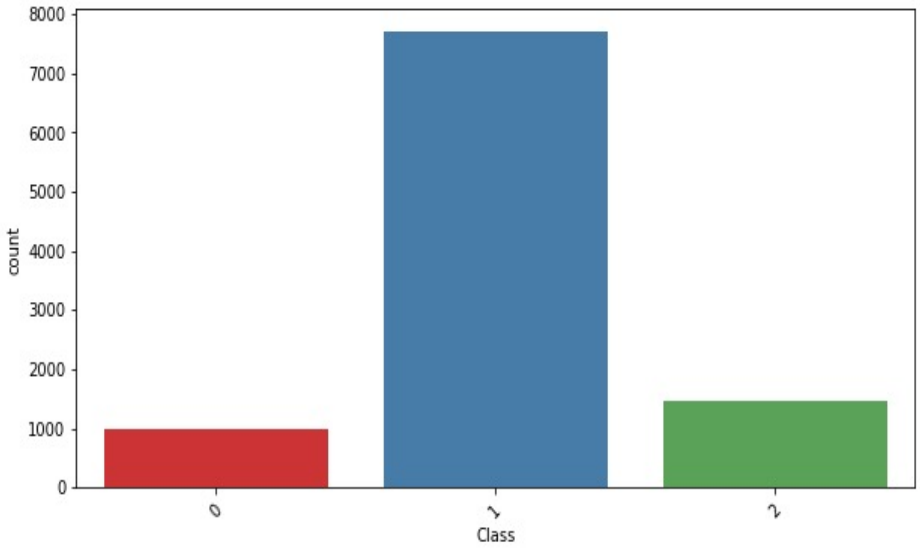

Figure 16: Multi Korattur data-set

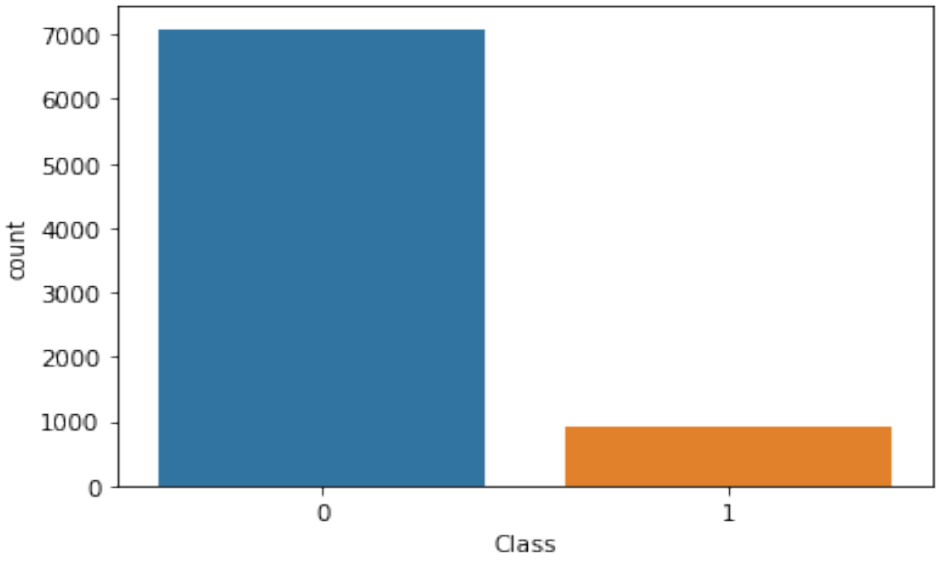

Figure 17: Binary Kaggle data-set

5.3 Distribution of the parameters

5.3.1 Boxplot

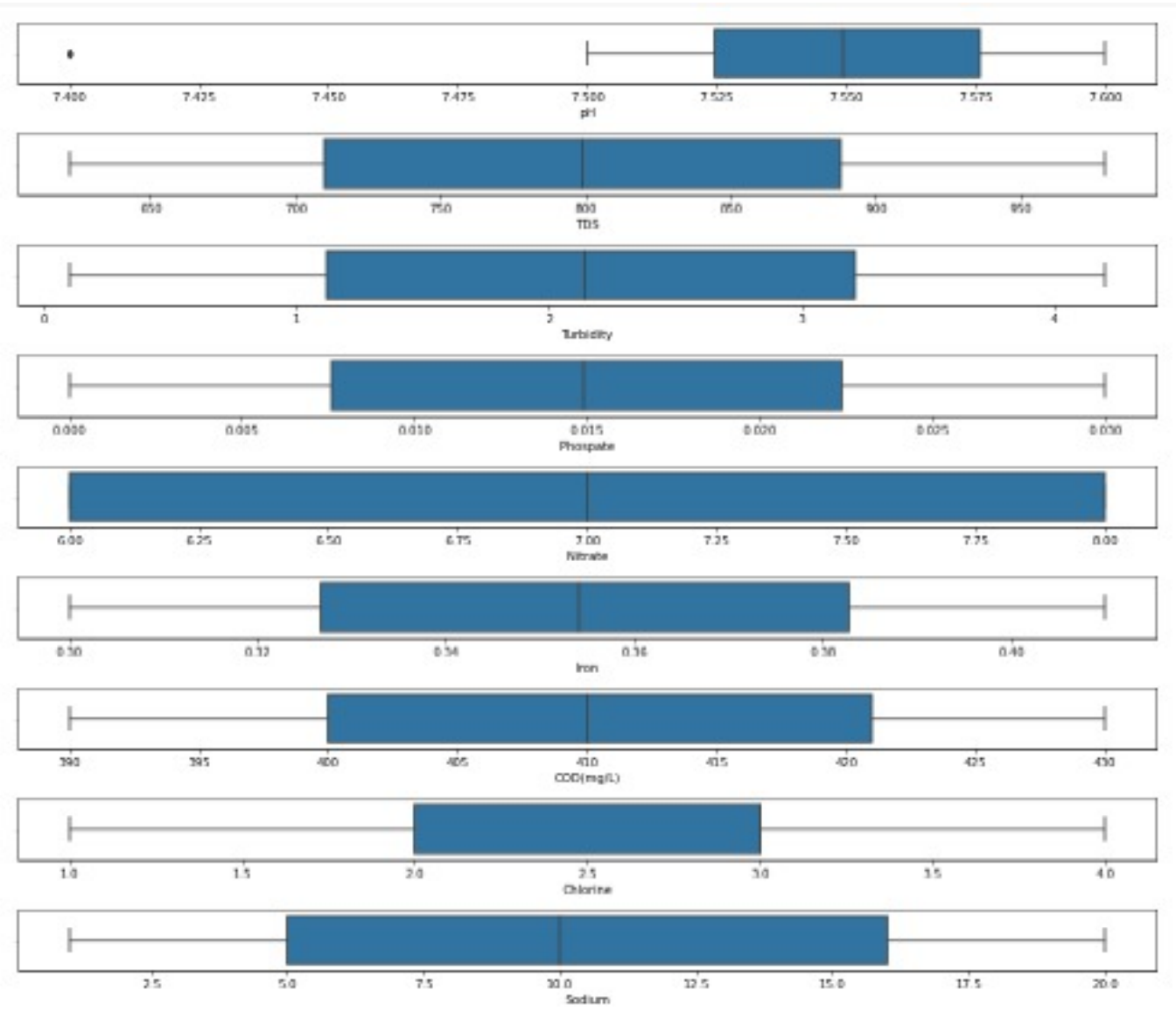

Figure 18: Binary Korattur data-set

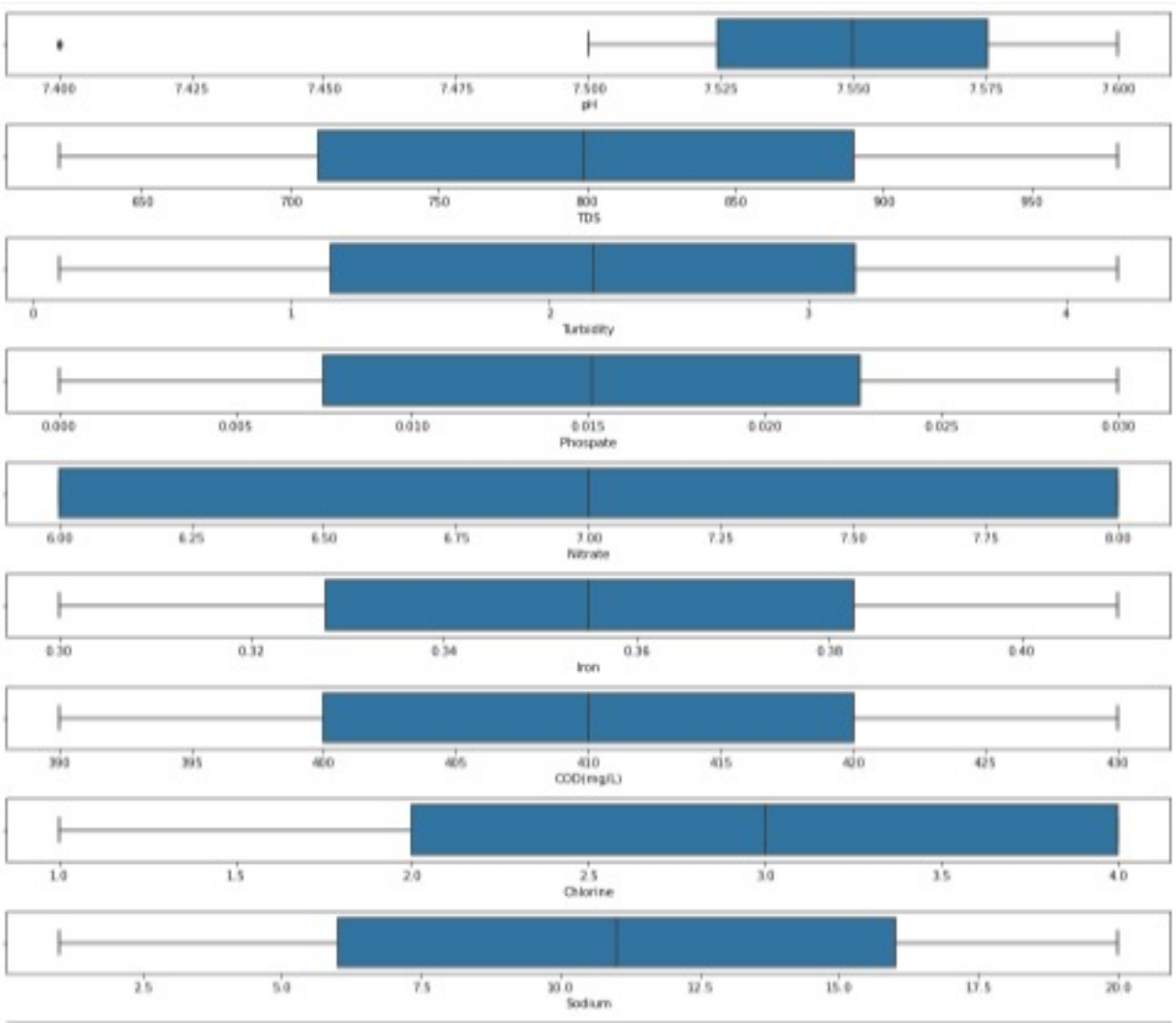

Figure 19: Multi Korattur data-set

5.3.2 Radar chart

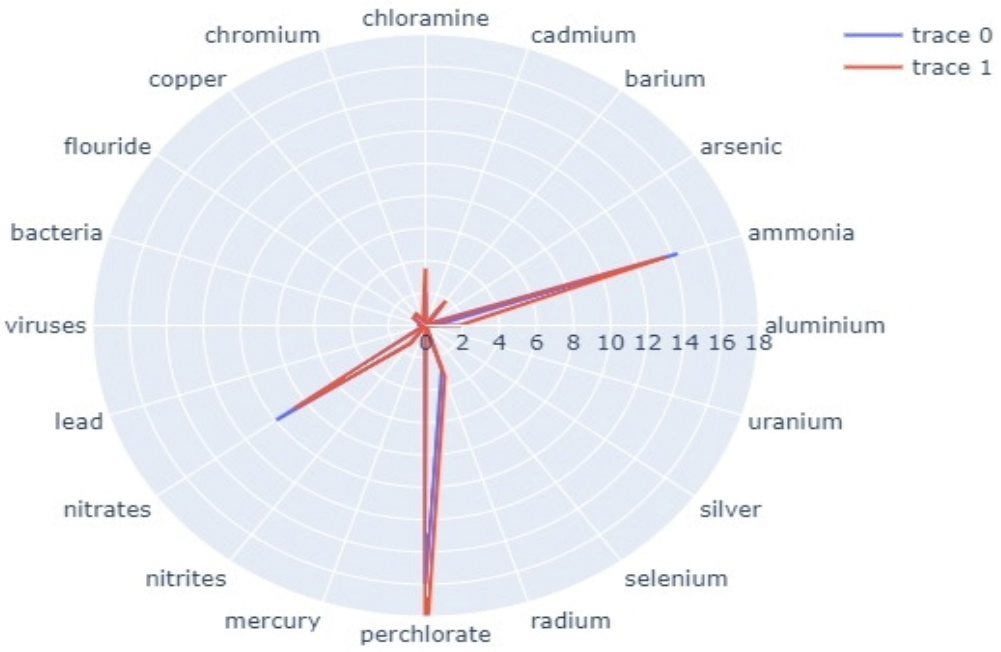

Figure 20: Binary Kaggle data-set - Overview from range 0 to 18

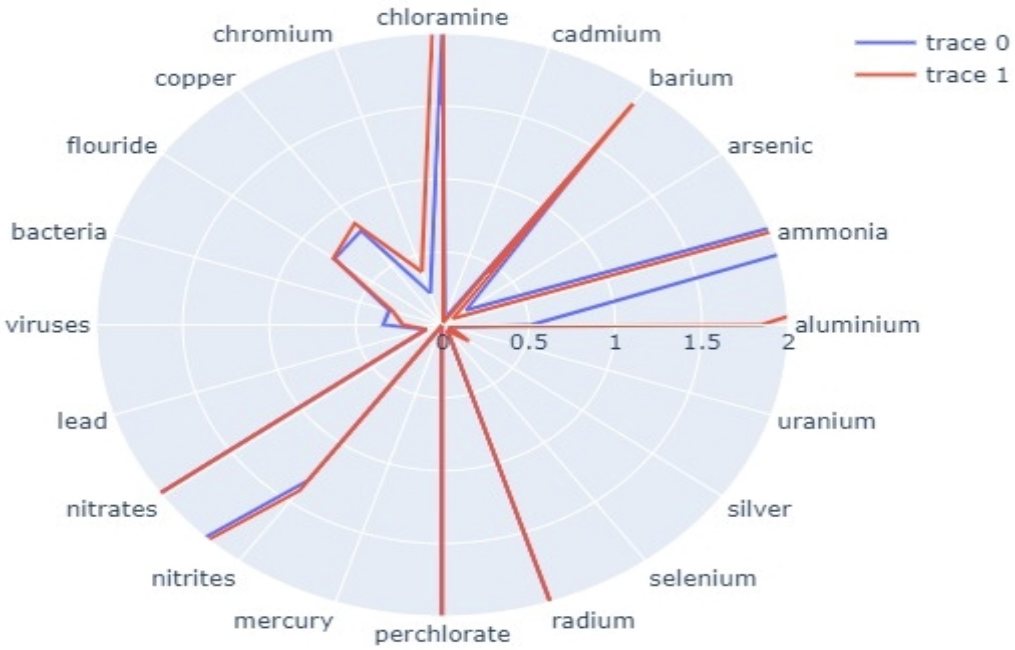

Figure 21: Binary Kaggle data-set - Overview from range 0 to 2

5.4 Heatmap

5.4.1 Binary Class Korattur Lake Dataset

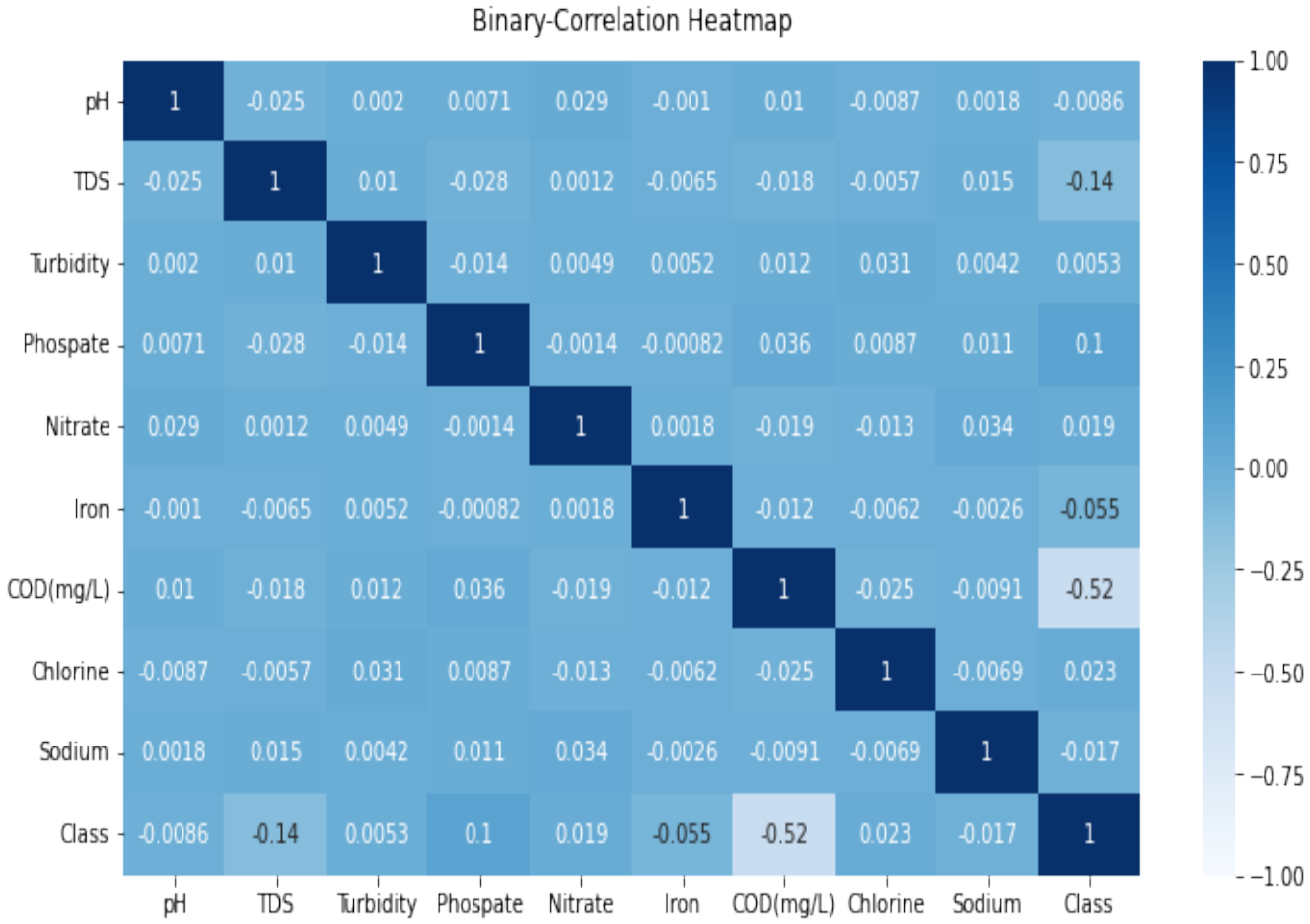

Figure 22:

5.4.2 Multi Class Korattur Lake Dataset

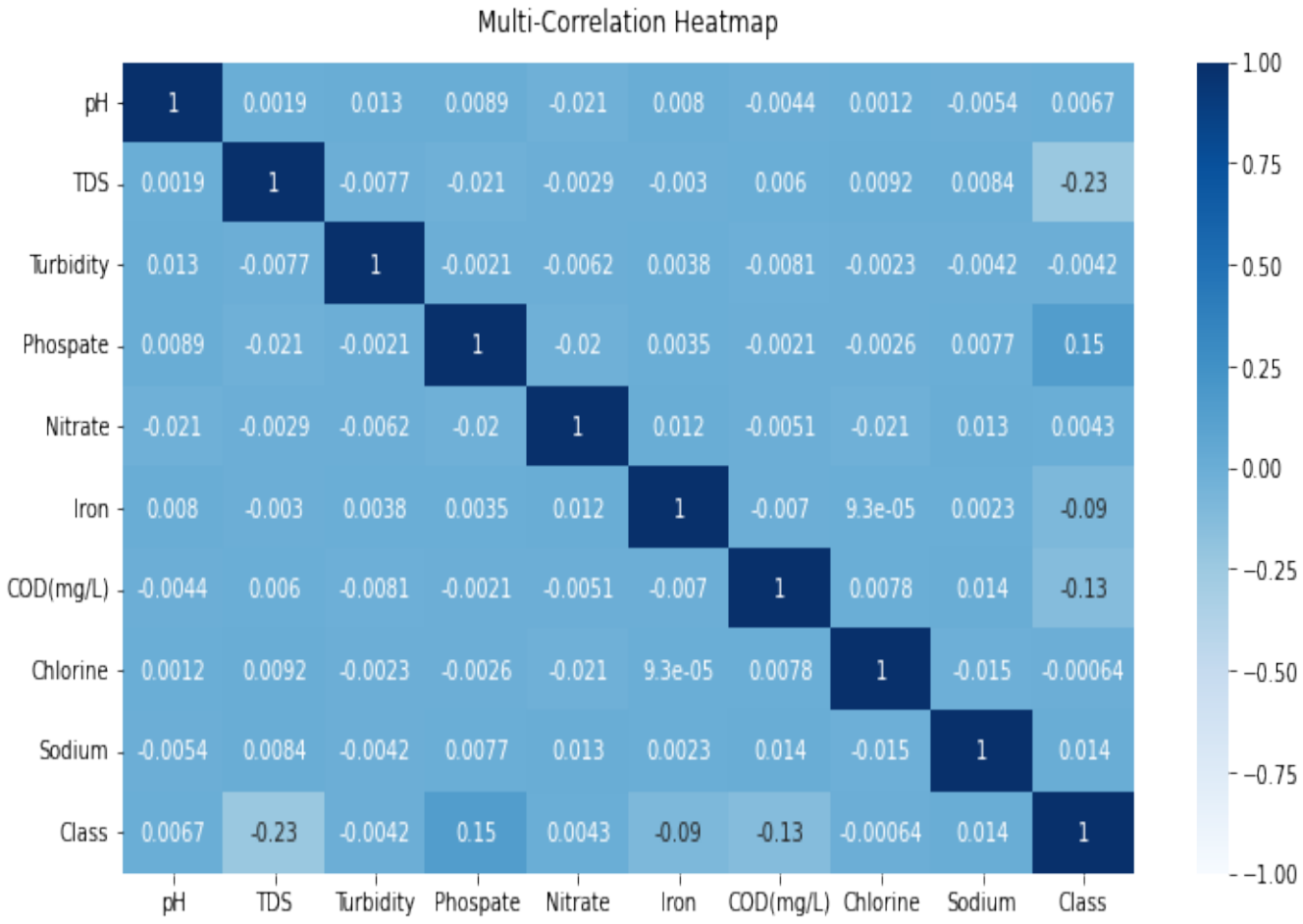

Figure 23:

5.4.3 Binary Class Kaggle Source Dataset

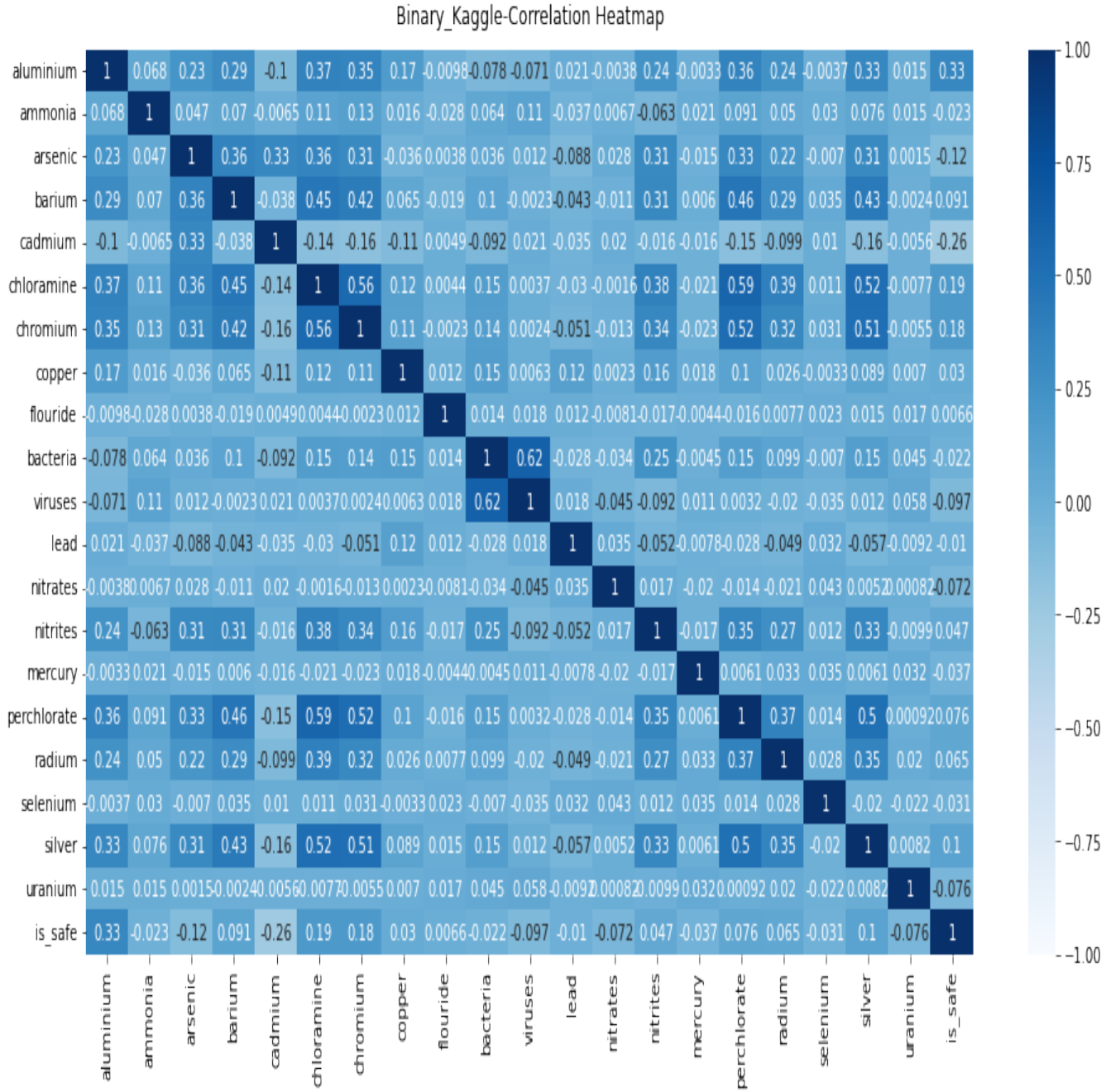

Figure 24:

6 Results

6.1 Outcomes

6.1.1 Accuracy-Korattur Lake Dataset- Binary class data

While analysing the accuracy of the models using Korattur Lake Dataset (Binary),the following outcomes were obtained. Analysing the dataset using statistical models, HCA has achieved an accuracy of 53%, followed by PCA which obtained an accuracy of 87%, LDA whose accuracy is 91% and QDA which achieved the highest accuracy of 95%. During the training of ensemble models both Stacking and Bagging achieved an accuracy of 99.8% while the highest accuracy was obtained by Boosting of 100%.

QDA	0.9519333333	**Bagging**	0.9988888889
LDA	0.9183826319	**Boosting**	1
HCA	0.5334	**Stacking**	0.9988888889
PCA	0.871		

Figure 25: Accuracy-Korattur Lake Dataset- Binary class data

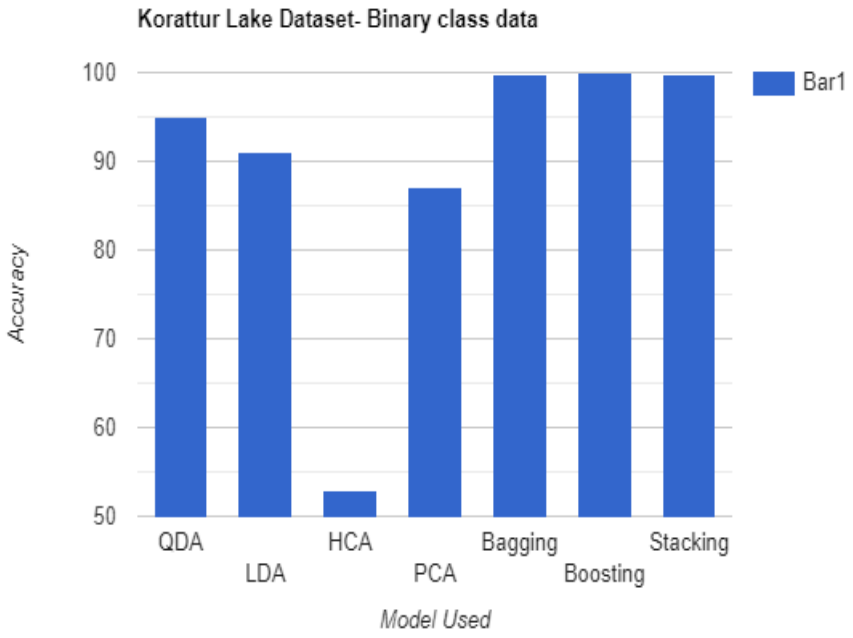

Figure 26: Barchart for Model and Accuracy

6.1.2 Accuracy-Korattur Lake Dataset- Multiclass data

While analysing the accuracy of the models using Korattur Lake Dataset (Multiclass),the following outcomes were obtained. In the pre-processing using statistical models, Quadratic discriminant analysis got a 95% accuracy, Linear Discriminant Analysis got a 92% accuracy, Hierarchical Clustering Analysis got a 39% accuracy and Principal component analysis got an 76% accuracy. When it comes to the ensemble models, Bagging, Boosting and Stacking got a 100% accuracy.

QDA	0.9497651503	**Bagging**	1
LDA	0.923589166	**Boosting**	1
HCA	0.3899792879	**Stacking**	1
PCA	0.758382643		

Figure 27: Accuracy-Korattur Lake Dataset- Multiclass data

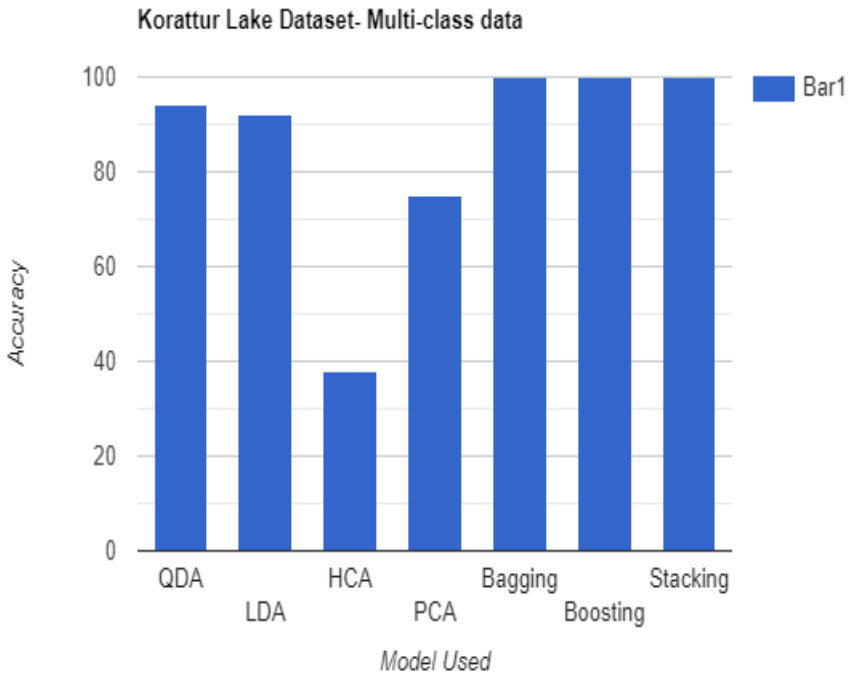

Figure 28: Barchart for Model and Accuracy

6.1.3 Accuracy-Kaggle Dataset- Binary class data

While analysing the accuracy of the models using Kaggle Dataset (Binary),the following outcomes were obtained. In the pre-processing using statistical models, Quadratic discriminant analysis got an 87% accuracy, Linear Discriminant Analysis got an 89% accuracy, Hierarchical Clustering Analysis got a 57% accuracy and Principal component analysis got an 89% accuracy. When it comes to the ensemble models, Bagging got a 96% accuracy, Boosting got an 89% accuracy and Stacking got a 96% accuracy.

QDA	0.8783440759	**Stacking**	0.965625
LDA	0.8967240822	**Bagging**	0.96796875
HCA	0.5735716965	**Boosting**	0.890625
PCA	0.889375		

Figure 29: Accuracy-Kaggle Dataset- Binary data

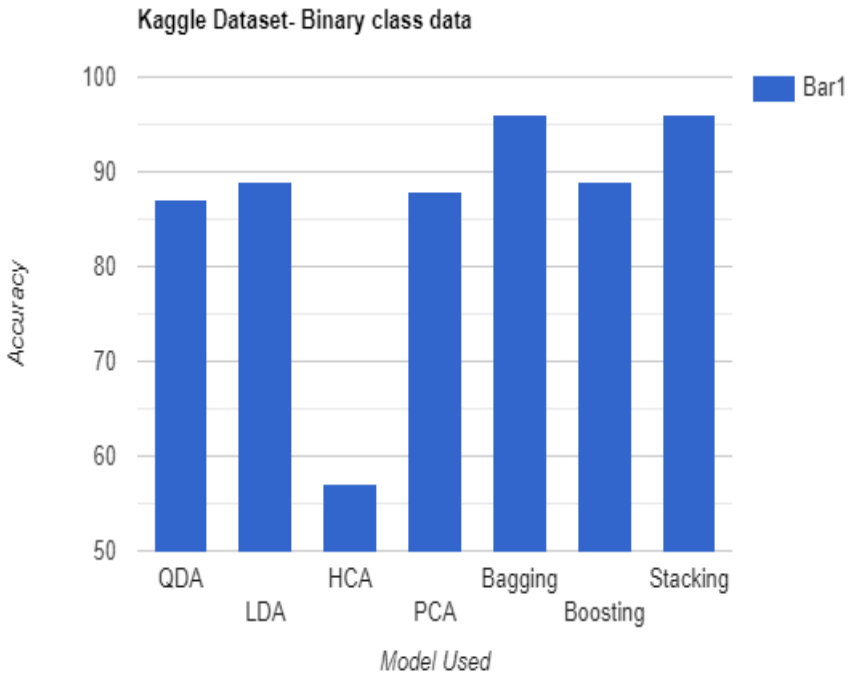

Figure 30: Barchart for Model and Accuracy

6.2 Performance Graphs

6.2.1 Binary Class Dataset - Korattur Lake

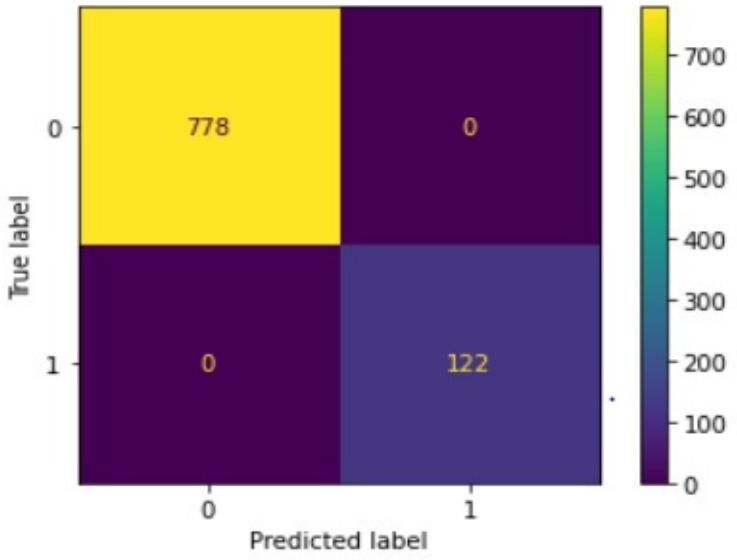

Figure 31: Confusion matrix

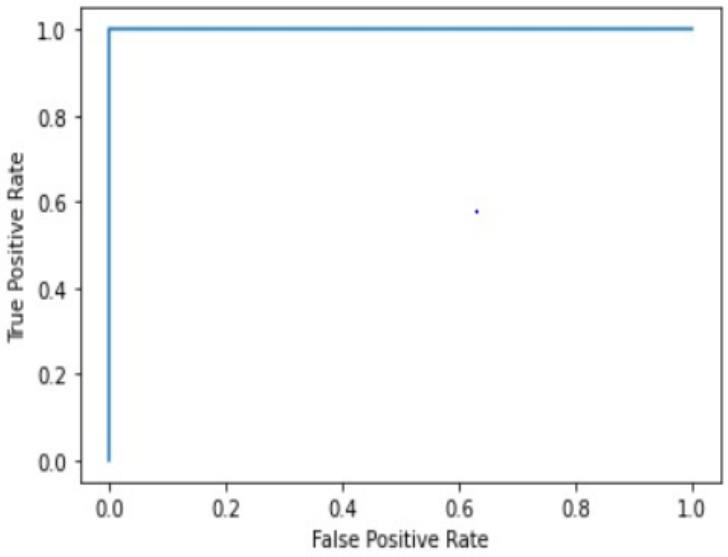

Figure 32: AUC-ROC curve

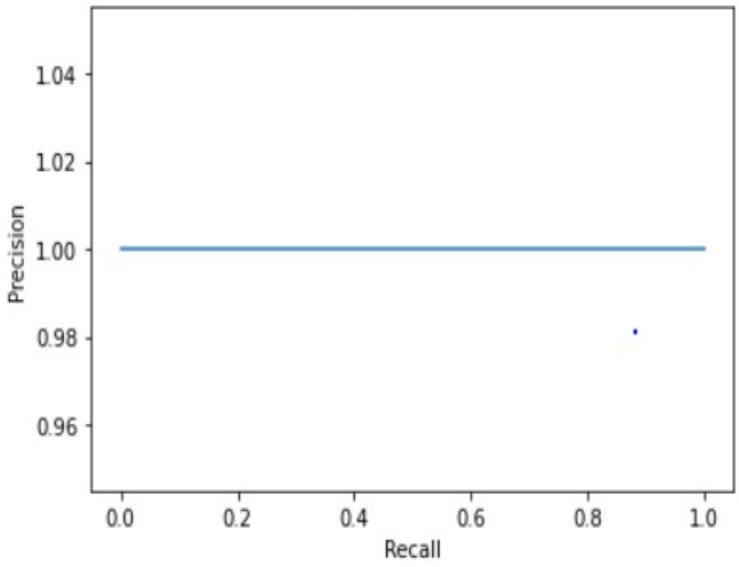

Figure 33: Precision-recall curve

6.2.2 Multi Class Dataset - Korattur Lake

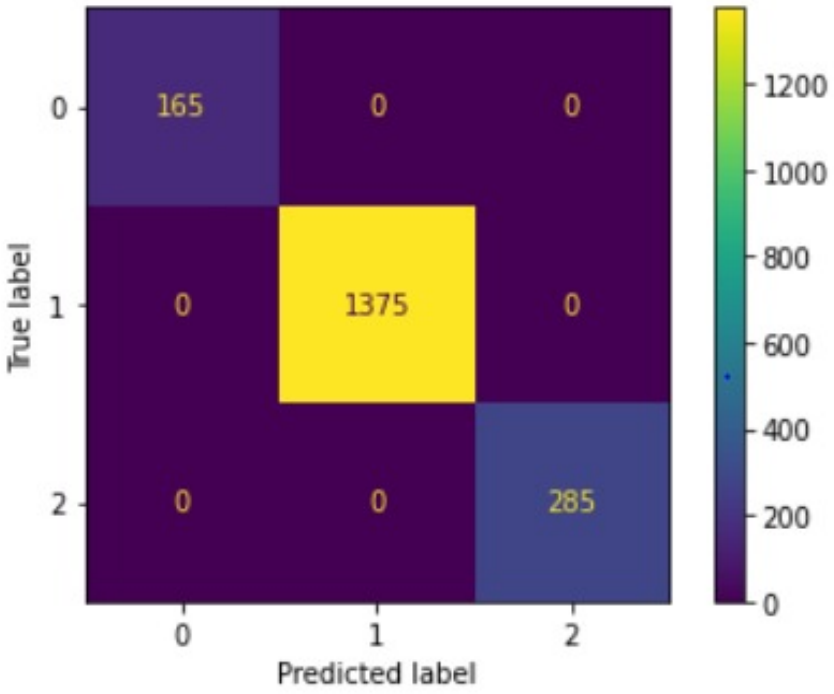

Figure 34: Confusion matrix

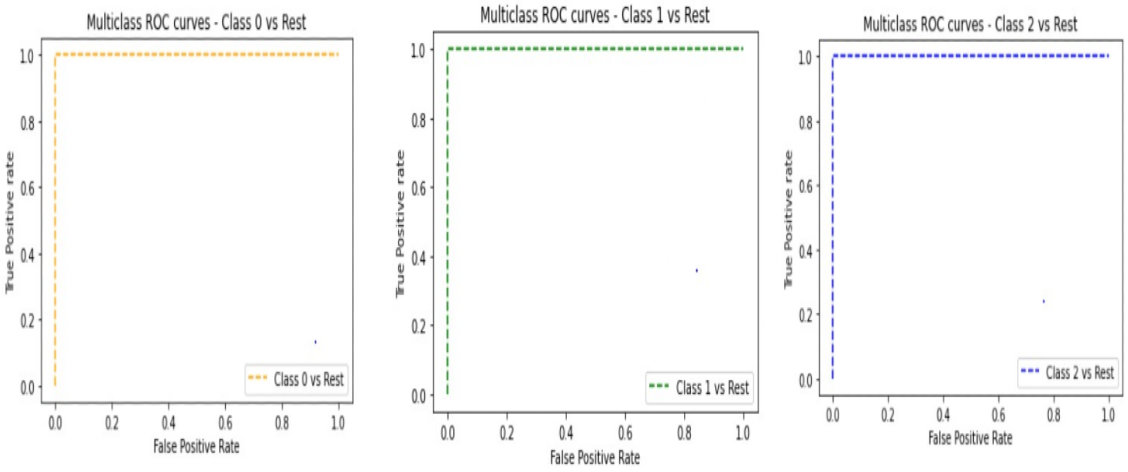

Figure 35: AUC-ROC curve

6.2.3 Binary Class Dataset - Kaggle

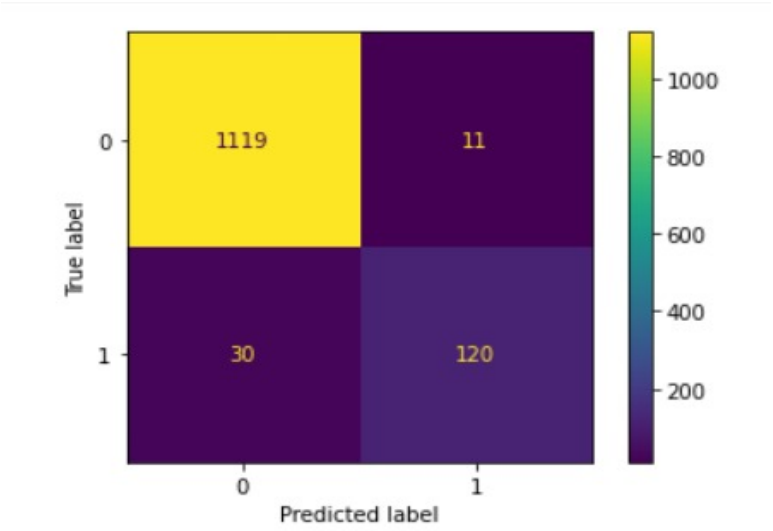

Figure 36: Confusion matrix

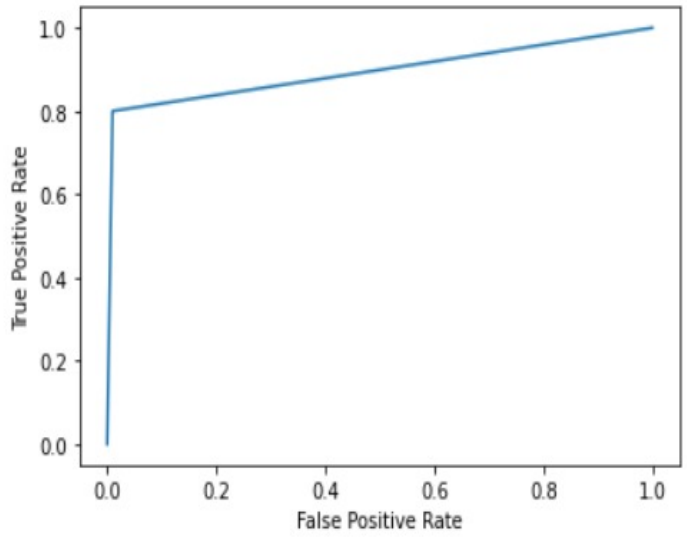

Figure 37: AUC-ROC curve

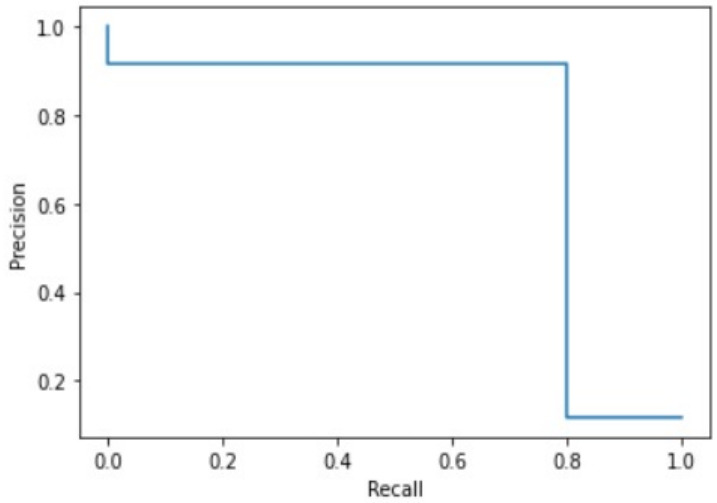

Figure 38: Precision-recall curve

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

[14] https://www.kaggle.com/datasets/mssmartypants/water-quality