**WATER QUALITY PREDICTION**

**TEAM MEMBERS:**

**2103A52012 - Bala Varshitha Duggempudi**

**2103A52029 - Tejaswi Prathap**

**1.ABSTRACT**

Every common man’s right, revolves around safe and potable water is unavailable to many regions of people. For purpose of assuring water portability and preserving public health, water source quality monitoring is essential. Innovative solutions are necessary since conventional methods for evaluating the quality of water can be time-consuming and expensive. This study investigates how machine learning (ML) models might improve the evaluation of water portability. The creation of ML-based models adapted to particular water predicting indicators, like turbidity, pH, dissolved oxygen, microbiological contamination, is significant contribution of this research. These models can make reliable predictions because they were trained on large datasets collected from various water sources. The research also discusses difficulties in scaling ML-based water portability assessment systems and issues with data quality, model interpretability, and scalability. In summary, this study highlights how machine learning can revolutionize the way that water quality is assessed for better water portability. The results offer useful information for researchers, environmental organizations, and legislators who want to use ML to get access in communities across the world obtain clean water. Finally, in a time of growing water scarcity and environmental challenges, this study advances sustainable and effective water management techniques. The final results shows that out of some machine learning models used in dataset (SVM, Random Forest, XGBoost, AdaBoost, logistic regression, KNN), SVM classifier has achieved highest accuracy.

**2. INTRODUCTION**

Water reusability can help in improving pure water, hygienity and improve water usage, only when cleaned safe for drinking.[1.1]

The quality of water targets relies on water reusability, various water predicting systems could help in treating waste water like household purpose, agriculture and industrial scales.[1.2]

On comparing power of bacterial and viral growth in water systems of the ML algorithms like logistic-regression, SVM , RF. These are trained by minimising the error, whether the water is safe to drink or not, and also we made a brief analysis for consequences of false predictions.[1.3]

SVMs are type of generalised non-linear algorithm for classification a continuous and categorical data ,help in solving complex data by optimisation. Using SVR and SVC methods, monitoring the groundwater is made easy.[11.2]

ANN , is easy to estimate the complicated behaviour . Additionally, this can optimise the scenarios with various conditions in a more efficient way. On comparison with many assumptions, complicated input values, parameters of models, measuring the input variables of algorithms are simplified. In contrary, ML helps in predicting that lack hydrogeological survey data. [11.3]

SVM is a linear classification algorithm based on the maximum distance. Support vector machine are the most popularly used ML algorithms for predicting as they help in maximizing the accuracy of predictions. SVM’s follow kernel functions like liner KF, polynomial KF, Gaussian KF and sigmoid KF, which has capability of optimising non-linear predictions of the algorithms. [11.3.1]

The abundance and complexity of the variables that define water quality make it difficult to measure and calculate water quality. Therefore, water quality indexes have been created to access the acceptability of water for various uses. This idea compares the appropriate standards and the water quality measure. Non-water professionals can easily understand the results of the WQI, which presents significant amount of quality data of water. [17.2]

A decision-tree is a technique for analysing the predictions. The decision-tree is a prediction algorithm used here. It is used for analysing observations about a node represented by branches to the nodes predicted value, represented by leaves. Decision Trees are popular ML approaches because of their readability and simplicity. The nodes in the decision-tree are divided into sub-nodes considering attribute’s threshold value. The CART algorithm uses the Gini index criterion to find the best homogeneity for the sub-nodes. The root node in dataset is used as train data, and the best feature and the threshold value to split into two parts. Additionally, the subsets are divided using the same rationale. This process is continuously occurred till the decision-tree has the least pure sub-set or the maximum no. of leaves in that growing tree. Tree pruning is another name for this. [13.3.2]

Decision Trees are popular ML approaches because of their readability and simplicity. The nodes in the decision-tree are divided into sub-nodes considering attribute’s value. Therefore, the water quality indices have been created to access the acceptability of water for various uses. This idea compares the appropriate standards and the water quality measure. Non-water professionals can easily understand the results of the WQI, which presents significant amount of quality data of water. [5]

Support Vector Machines are a ML techniques based on statistical learning theory. On estimate of dependencies and predictive learning from final data sets, VC-theory has a solid mathematical foundation. SVM is dependent on structural risk minimisation principle, which aims to reduce both empirical risk and model complexity while maintaining good generalizability. SVM is a supervised learning algorithm that is used to solve classification and regression models such as SVC and SVR. However, it is only used on small datasets because computing them takes too long. [13.3.3]

Supervised Vector Machines are ML algorithms used in continuous and categorical data. SVMs are more dominant than regression models, but they work best with limited datasets. First, every data variable is plotted in an n-dimensional state, with n equalling the no. of characteristics. A hyper-plane creates to divide that may be to classify or to group the clusters physically. This approach uses the hyper-plane to increase the distance among classes on ignoring outliers. When linear separation is not achievable, kernels alter data to make it more separable. [13.3.4]

**3.LITERATURE SURVEY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **AUTHOR &**  **YEAR** | **MODEL USED** | **MERITS** | **DEMERITS** | **FUTURE SCOPE** | **PARAMETERS** |
| 1.Eva Reynaert (2023) | Logistic Regression  SVM  Random Forest. | Better interpretability  Well-suited for early warning systems . | Standard operation data.  Highly complex or rapidly changing systems. | Enhance prediction accuracy and reduce sensor costs. | Logistic regression  True Positive Rate (TPR). |
| 2.Osim Kumar Pal  (2022) | RF  SVM  ANN  DNN  Gaussian Naive Bayes. | A quicker and potentially less costly alternative to conventional laboratory techniques. | Does not explicitly mention specific drawbacks | Exploration of IoT-based quality detection models . | Hardness  Conductivity  pH |
| 3.Mohammad Reza Goodarzi  (2023) | MARS for WQI. | Machine Learning Applications  Comparative Analysis | Limited Model Selection | Algorithm Exploration  Real-time Monitoring  Climate Change Impact | ph, turbidity  DO  TDS  Nitrate levels. |
| 4.Xiaotong Cen(2023) | CNN  ANN  Logistic regression. | Comprehensive Review  Highlighting Research Directions. | The abstract does not detail any specific models. | Real-time Control and Optimization.  Integrated Application. | chemical dosages  effectiveness of different  chemicals. |
| 5.Darío Calzadilla Cabrera  (2023) | SVM  Decision tree  KNN. | Highlights potential treatment methods for MP pollution control. | Absence of such details might limit the methods | Encouraging the use of natural textiles. | Sludge samples  Vegetation density  Wetland design |
| 6. Hamza Ahmad Madni (2023) | Stacked ensemble H2O  KNN | Accuracy Improvement  Handling Missing Values. | Does not explore the limitations | Explore a broader range of models including deep learning architectures | Accuracy of predictions  Precision  F1-score. |
| 7.Mohammad RezaGoodarzi  (2023) | GEP  M5P Model Tree  MARS | Comprehensive Water Quality Assessment. | Limited Model Selection  Model Performance Interpretation | Predict WQI in real-time by incorporating live data feeds. | RMSE  MAE |
| 8.[Amir Hamzeh Haghiabi](javascript:;)  (2018) | ANN  GMDH  SVM | Explores the impact of different transfer functions. | Potentially limiting the understanding of the best-performing model. | Potentially achieve even better prediction accuracy. | pH, SO4, Na, Ca, Cl, Mg,  HCO3  Prediction accuracy metrics  Transfer functions  Kernel functions. |
| 9. M Vijay Anand(2023) | CNN with Tensorflow  and Keras. | CNN, Tensorflow, and Keras to predict water quality based on image analysis. | It does not consider other water quality parameters. | Hybrid models that combine image analysis could lead to even better results. | CNN architecture  (Tensorflow, Keras)  Image analysis,  Image color  Image quality. |
| 10. [J. I. Ubah](https://www.nature.com/articles/s41598-021-04062-5#auth-J__I_-Ubah-Aff1)(2021) | ANN  TDS  EC | ANN to predict water quality index. | Focuses on only four parameters pH, TDS, EC, Na. | Enhance the accuracy and reliability. | pH,TDS  EC,Sodium  RMSE. |
| 11. Zhenjiang Wu, Chuiyu Lu(2023) | SVM  LSTM  MLP  GRU | Model performed dynamically | Data segmentation  Subsequence length. | Impact of different activation functions provide valuable insights. | RMSE  R²  NSE |
| 12.Jaehoon Kim , HyeonseopYuk  (2023) | SVM  Random Forest | More efficient  High throughput solution. | Lack of GPU hardware. | Intend to develop system with GPU hardware. | Faster R-CNN  YOLO  SSD  FairMOT |
| 13.Jagadish Kumar Mogaraju  (2023) | Neural Ntworks | Interpolation methods for groundwater prediction. | Did not mention any limitations | Population induced pressure is highlighted. | No such parameters. |
| 14. Tianlong Jia , Zoran Kapelan(2023) | Computer Vision. | Automatic detection of microplastic litter. | Lack of open access to code. | Lack of open access to codeTop of Form | sampling and visual  observations. |
| 15. Khabat Khosravi ,  Fatemeh  Rezaie (2023) | CNN  RNN  LSTM | SWE susceptibility using satellite data, rainfall, soil. | No specific limitations. | Monitoring data is lacking. | CNN  RNN  LSTM |
| 16. Nida Nasir, Afreen Kansal  (2023) | Stacked Ensemble Model. | Achieves accuracy on the dataset. | Doesnot elaborate specific algorithms. | Alternate ensemble methods could evaluate the model’s effectiveness. | Decision Tree Classifier. |
| 17.Mustafa YURTSEVER, Murat EMEC  (2023) | SXH Hybrid Model. | Significant improvement over existing models. | Lacks details of algorithms used. | Further research may enhance its interpretability. | Achieves better accuracy  and F1-Score. |
| 18. Aissam Gaagai, Hani Amir Aouissi  (2023) | ANN  GBR | Utilises machine learning techniques for groundwater quality assessment. | Specific details of algorithms are not mentioned. | More access to real world data could improve the efficiency. | ANN  GBR  ANN-2F |
| 19. Peng Chen , Biao Wang (2023) | Self-optimising machine learning algorithm. | Utilises multi sensing data. | Insufficient data and accuracy levels. | Improving accuracy and expanding the range of monitoring. | Improve quality water  prediction. |
| 20. Md Galal Uddin , Stephen Nash  (2023) | SVM  RF | More efficient and high throughput. | No specific algorithms are mentioned. | Enhance the accuracy and reliability. | SVM  RF |

3.1 SUMMARY

Eva Reynaert’s 2023 [1] paper explores ML methods- logistic regression, SVM, and RF. It suggests hybrid models combining system knowledge with machine learning to enhance prediction accuracy and reduce sensor costs, particularly utilising logistic regression and true positive rate (TPR).

Osim Kumar Pal’s 2023 [2] paper investigates on RF, SVM, ANN, DNN, Gaussian Naïve Bayes. The focus is on predicting and analysing drinkable water quality across regions. Parameters like hardness, conductivity, pH, organic carbon, and solids are considered.

Mohammad Reza Goodarzi’s 2023[3] paper focusses on employing multivariate adaptive regression splines (MARS) for water quality index (WQI) prediction. Parameters like pH, turbidity, dissolved oxygen, nitrate levels are considered.

Xiaotong Cen’s 2023 [4] paper offers discussing CNN, ANN and logistic regression models. It provides a comprehensive assessment with focus on sustainability and emerging research directions. The review emphasises real-time control, optimization aiming for comprehensive studies. It covers aspects such as chemical dosages, effectiveness of various chemicals, economic costs, energy consumption, environmental impacts and sewer system dynamics.

Darío Calzadilla Cabrera’s 2023 [5] paper utilises SVM, decision tree, KNN models. It examines the efficiency of treating wetlands in removing microplastics from waste water, offering potential methods for pollution control. The research considers influent and effluent MP concentrations, sludge samples, vegetation density.

Hamza Ahmad Madni’s 2023 [6] paper introduces using a stack ensemble model from H2O auto ml for accurate water-quality prediction, along with KNN imputer to manage missing values. The focus is on enhancing accuracy. Suggests broader model exploration, possibly including deep learning.

Mohammad RezaGoodarzi’s 2023[7] paper focusses on utilization of Gene Expression Programming (GEP), M5P mode tree, MARS for predicting water quality index (WQI). It aims for a comprehensive assessment and model comparison, and performance interpretation might pose challenges. Evaluation metrics include RMSE, MAE, NSE, along with water quality parameters as dataset features.

[Amir Hamzeh Haghiabi](javascript:;)’s 2018 [8] paper conducts employing ANN, GHMD, SVM. It assesses the influence of transfer functions for ANN and kernel functions for SVM on prediction performance. Parameters includes pH, SO4, Na, HCO3, prediction accuracy metrics, error indexes.

M Vijay Anand’s 2023 [9] research utilises CNN with Tensorflow and keras. The model predicts water quality through image analysis. The study focus excludes other parameters like pH, turbidity, chemical composition. The analysis involves CNN architecture, tensorflow, keras, water quality prediction accuracy, image color.

[J. I. Ubah](https://www.nature.com/articles/s41598-021-04062-5#auth-J__I_-Ubah-Aff1)’s 2021 [10] study investigates ANN. It predicts water quality based on pH, TDS, EC, and Na. Incorporating broader range of water quality parameters in ANN model could enhance accuracy. The paper evaluates pH, TDS, EC, Na R2 values, RMSE and water quality index.

Zhenjiang Wu, Chuiyu Lu’s 2023 [11] paper investigates on MLP, GRU, SVM models, the main ones functional in research. For varying and rising stations, the GRU model outperforms, with simulation discrepancies related to segmentation, subsequence length, and parameter uncertainty. Metrics for evaluation include RMSE, R2, and NSE.

Jaehoon Kim , HyeonseopYuk’s 2023 [12] research describes assessing water toxicity using the locomotor responses of Daphnia Magna. It seeks to deliver a quicker high-throughput solution by utilizing RF and SVM. Due to GPU hardware restrictions, Faster R-CNN, YOLO, SSD, and DETR were not effective.

Jagadish Kumar Mogaraju’s 2023 [13] project uses neural networks to do research in India's YSR district. In order to understand groundwater dynamics and alleviate stresses connected to population growth, the research emphasizes potentiality of DL models, NN.

Tianlong Jia , Zoran Kapelan’s2023 [14] review investigates the use of computer vision .It emphasizes the potential of deep learning for automatic macro-plastic litter detection. The report emphasizes the need to fill knowledge gaps in riverine ecosystems, which are essential for the transport and storage of litter.

Khabat Khosravi , Fatemeh Rezaie’s 2023 [15] paper, we emphasize CNN, RNN, and LSTM. To determine the susceptibility of soil water erosion, deep learning models are used in conjunction with satellite data, rainfall data, and soil data. The study examines CNN, RNN, and LSTM, emphasizing their effectiveness and suitability in this situation.

Nida Nasir, Afreen Kansal’s 2023 [16] research uses a Stacked Ensemble Model, which consists of five stacked neural networks plus a machine learning classification models. The study's astounding accuracy on the dataset demonstrates its superiority to earlier research.

Mustafa YURTSEVER, Murat EMEC’s 2023 [17] research focuses on using a hybrid model (SXH) that outperforms existing algorithms with regards to accuracy , F1-score. Model's interpretability and reliability could be improved by further study by elaborating on predictions.

Aissam Gaagai, Hani Amir Aouissi’s 2023 [18] study make use of Gradient Boosting Regression (GBR) and ANN algorithm. For a thorough evaluation of groundwater quality, the study incorporates a number of tools, including IWQIs, multivariate analysis, GIS, and machine learning. However, ANN and GBR models' repeatability and wider applicability are hampered by the absence of specifics.

Peng Chen , Biao Wang’s 2023 [19] paper, In order to solve sample size and scale irregularities, the study keeps in track the water quality of urban rivers. The study is aware of its limitations, including the lack of sufficient data and accuracy issues. For more thorough insights, improving accuracy and extending monitoring range are proposed.

Md Galal Uddin , Stephen Nash’s 2023 [20] study describes use of machine learning techniques like Gaussian Process Regression (GPR) and Monte Carlo Simulation (MCS). The work deals with the important problem of modelling uncertainty in WQI, which is crucial for accurate water quality evaluation and management.

3.2 PROBLEM STATEMENT

* Develop microbial water quality monitoring models for MBR+Cl2 water reuse systems.
* Prioritize client security by minimizing false positive predictions while providing early warnings for unsafe water conditions.
* Customize algorithms to specific water predictions according to various re-usable applications.

**4. METHODOLOGY**

4.1 ARCHITECTURE:

Feature Selection

Data Cleaning

Data Preprocessing

Water Quality dataset

Model Optimisation

Model Initialization

Model Training with cross validation

Feature Engineering

Evaluation of Accuracy Using ML Models

SVM, LR ,RF , DT, KNN , XG Boost, AdaBoost

4.2 STEPS:

1. Dataset Selection
2. Data Preprocessing and Cleaning
3. Splitting data to train and test
4. Selecting Feature from the dataset
5. Model Initialization
6. Model Training with Cross Validation
7. Generate data and take new data
8. Model Optimization
9. Apply Feature Engineering
10. Apply ML algorithms on testing data
11. Evaluation of Accuracy

**5. EXPERIMENTAL WORK**

Set-Up Infrastructure:

* Google colaboratory
* Import required libraries Sklearn, Numpy, Pandas, Matplotlib ,Seaborn…

Data cleaning:

* Upload the collected dataset.
* Process the data into dataframe using Pandas library.
* Clean the null values if present in provided dataset.

Model Evaluation:

* Implement necessary ML models to the dataset.
* Evaluate the algorithm which achieves the more accuracy out of all ml algorithms.
* Take down the graphs and final result achieved.
  1. MATHEMATICAL FORMULA

*1.SVM*= [1/n

𝑛

∑ max(0,1) − 𝑦𝑖(𝑤𝑇𝑥𝑖 − 𝑏))] + 𝜆 ||w||^2

𝑖=1

Where ‘w’ – average vector , ‘xi’ – p dimensional real vector, ‘b’ – boundary

*2.LOGISTIC REGRESSION:*

π(X)=exp(β0+β1X1+… +βkXk)1+exp(β0+β1X1+… +βkXk)=exp(Xβ)1+exp(Xβ)=11+exp(−Xβ), π ( X ) = exp ⁡

*3.RANDOM FOREST:*

RFfi sub(x)= feature x is calculated from RF algorithm.

*4.DECISION TREE:*

Decision tree formula for classification:

I\_gini = 1 - Σ(p\_x^2)

Here:

I\_gini is the Gini impurity.

p\_x is the proportion of samples in node from class x.

Decision tree formula for regression :

MSE = Σ(y\_x - ŷ)^2 / a

Here:

MSE - Mean Square Error.

y\_x represents actual target values.

ŷ - predicted target.

a - No of samples.

*5.KNN:*

KNN for classification :

Y\_pred = argmax(Σ I(y\_b = c) for i in k-nearest neighbours)

Here: Y\_pred - predicted label for X.

y\_b - class label of knn.

c represents each unique class label in the dataset.

I(y\_i = c) is an indicator function that is 1 if y\_i equals the class label c and 0 otherwise.

KNN for regression :

Y\_pred = (Σ y\_b for i in k-nearest neighbours) / k

Where: Y\_pred - target value for X.

y\_b - target values of knn.

k - number of nearest neighbours.

*6.ADA BOOST:*

*Weighted Error (ε\_t) for Weak Classifier at Iteration t:*

ε\_t = Σ(w\_a \* I(y\_a ≠ h\_t(x\_a))) / Σ(w\_a)

Here:

ε\_t - weighted error at iteration t.

w\_a - weight assigned to x\_a.

y\_a - label of x\_a.

h\_t(x\_a) - prediction for training instance x\_a.

I(y\_a ≠ h\_t(x\_a)) – if weak classifier is wrong, then 1 else 0.

*Strong Classifier Weight (α\_t) for Weak Classifier at Iteration t:*

α\_i= 0.5 \* ln((1 - ε\_i) / ε\_i)

Here:

α\_i - value assigned at i.

ε\_i - weighted error at i.

*Update Weights for Training Instances at Iteration t+1:*

w\_i, t+1 = w\_i, t \* exp(-α\_t \* y\_i \* h\_t(x\_i)) for all i

Here:

w\_i, t+1 is the updated weight for t+1.

α\_t – value given to the weak classifier at iteration t.

y\_i - label for x\_i.

h\_t(x\_i) - prediction of x\_i.

*Final Prediction (Strong Classifier):*

H(x) = sign(Σ(α\_t \* h\_t(x)))

Here:

H(x) - final predicted value of x.

α\_t - value assigned at t.

h\_t(x) - prediction of x.

*7. XG BOOST:*

F(x) = Σ w\_i \* h\_i(x)

Here:

F(x) – prediction for x.

w\_i - value associated with the i-th tree.

h\_i(x) – predicted value of x.

5.2 DATASET DESCRIPTION

LINK:[**https://github.com/balavarshitha/Water-quality-prediction**](https://github.com/balavarshitha/Water-quality-prediction)

The dataset consists of 18,153 rows and 10 columns.

***pH value:*** pH plays a major factor in assessing balance of water. In addition, it also predicts the nature of water whether the water is acidic or basic. According to WHO, maximum pH values lies between 6.5 and 8.5, while current range lies between 6.52 and 6.83.

***Hardness:*** Hardness of water is determined by amount of calcium and magnesium present in salts. The water that pass through many geologic occurences releases such harmful salts. The hardness of water is determined by how much long it gets exposed in substance influencing in its raw state..

***Solids (Total dissolved solids - TDS):*** Numerous in-organic, some salts, like k, Ca, Na, bi-carbonates, Cl, Mg, sulphates can be dissolved in water. These minerals give the water an unpleasant flavour and nasty colour. This is a life-changing variable while using water. Water with a high TDS has a high mineral content. The recommended TDS level for drinking purposes is 500 mg/l, with a maximum limit of 1000 mg/l.

***Chloramines:*** Chlorine and chloramine are the major disinfectants used in many public water systems. Chloramines are formed when ammonia is mixed with chlorine for treating drinking water. Chlorine levels up to 4 mg/L or 4 ppm, are considered potable water.

*Sulphate:* These are organic compounds which naturally exist in rocks, soil, and minerals. They also exist in and as the surrounding air, groundwater, vegetation. Sulphate is mostly used for commercial purposes in the chemical industry. The amount of sulphate in saltwater is around 2,700 mg/L. The majority of freshwater sources have values between 3 and 30 mg/L, while certain regions have subsequently higher levels (1000 mg/L).

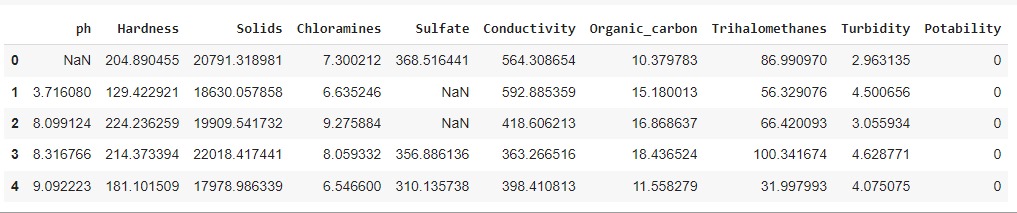
*Conductivity:* Pure water is a good insulator rather than a good conductive for electrical current. The electrical conductivity of water is improved by an increase in the ion concentration. The electrical conductivity of water is typically determined by the amount of dissolved particles. The ability of a solution to convey current through its ionic process is measured by electrical conductivity (EC). According to WHO guidelines, the EC value shouldn't be more than 400 S/cm.

***Organic\_carbon:*** On decomposing NOM and synthetic sources both contribute to TOC in the water sources. The TOC in organic compounds in pure water is a measurement of this. It estimates drinking water has 2 mg/L of TOC and that water source, which is used for treatment, contains 4 mg/Lit.

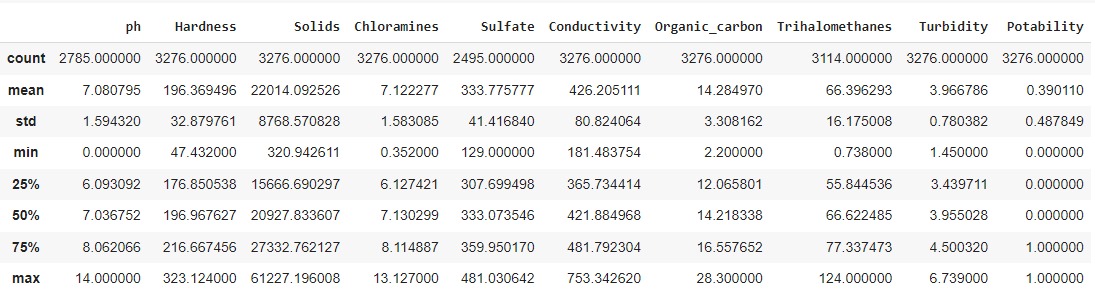
***Trihalomethanes:*** Chemicals called THMs can be discovered in chlorine-treated water. The amount of organic matter in the water, the quantity of chlorine needed to treat the water, and the temperature of the treated water, affect the concentration of THMs in drinking water. THM concentrations up to 80 ppm are regarded as safe for drinking water.

***Turbidity:*** The amount of solid stuff present in the suspended state determines how turbid the water is. The test is used to determine the quality of waste discharge with regard to colloidal matter and measures the light-emitting capabilities of water. The mean turbidity value (0.98 NTU) is less than the WHO-recommended threshold of 5.00 NTU.

***Potability:*** determines if water is drinkable or not, 1 - drinkable and 0 – undrinkable .



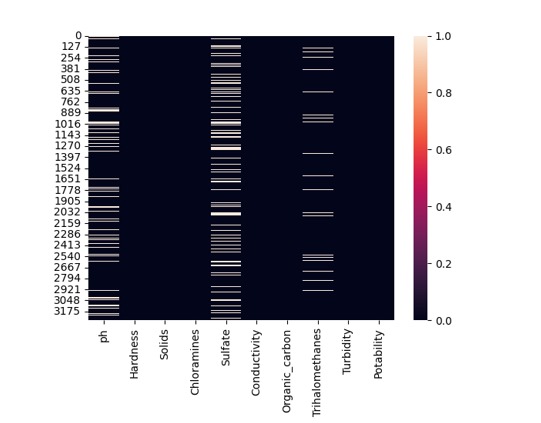
First 5 rows of dataset

****

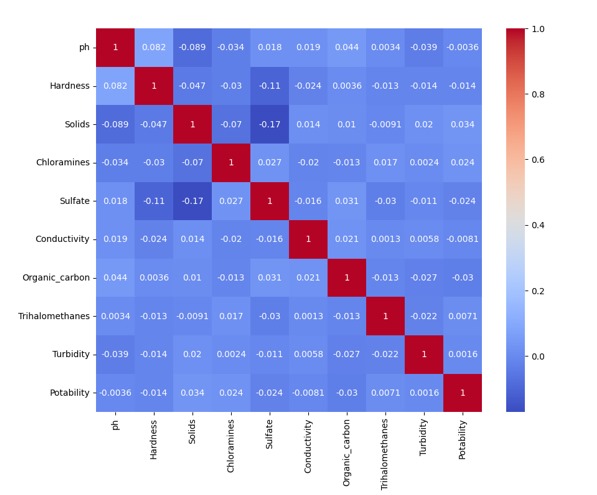
Describing the dataset

**6. RESULTS**

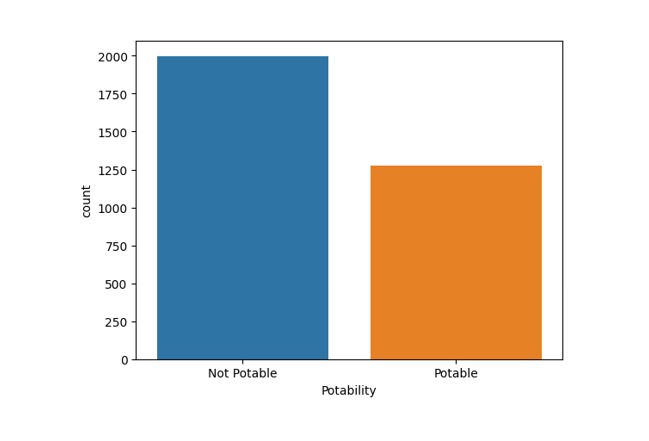
HEAT MAP:Graphical representation of heterogeneous data which are formatted as rows and columns in matrix. They help in evaluate co-relation over various numerical variables, analysing patterns.

****

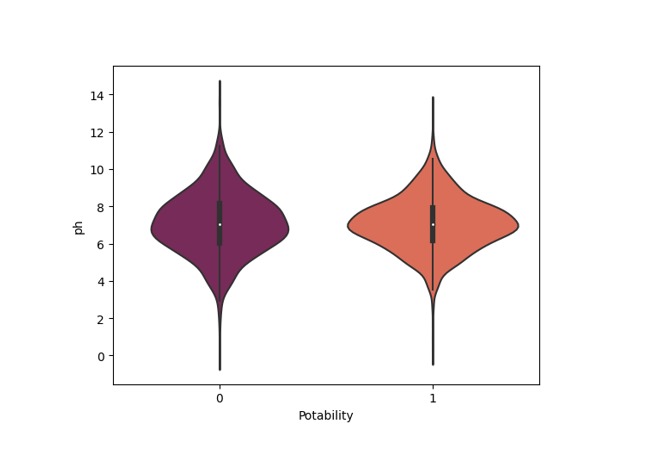
Heat map

****

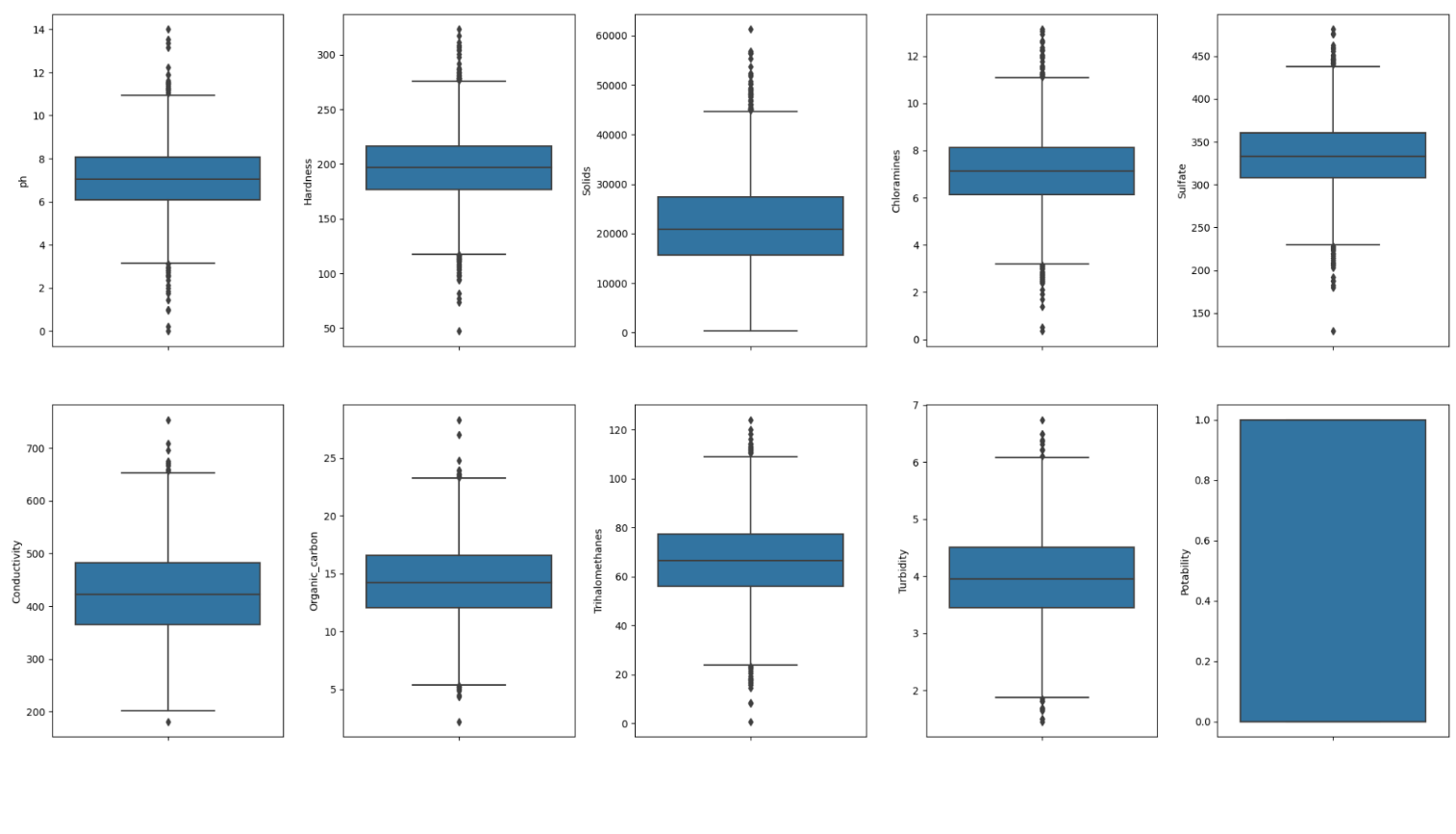
Correlation matrix



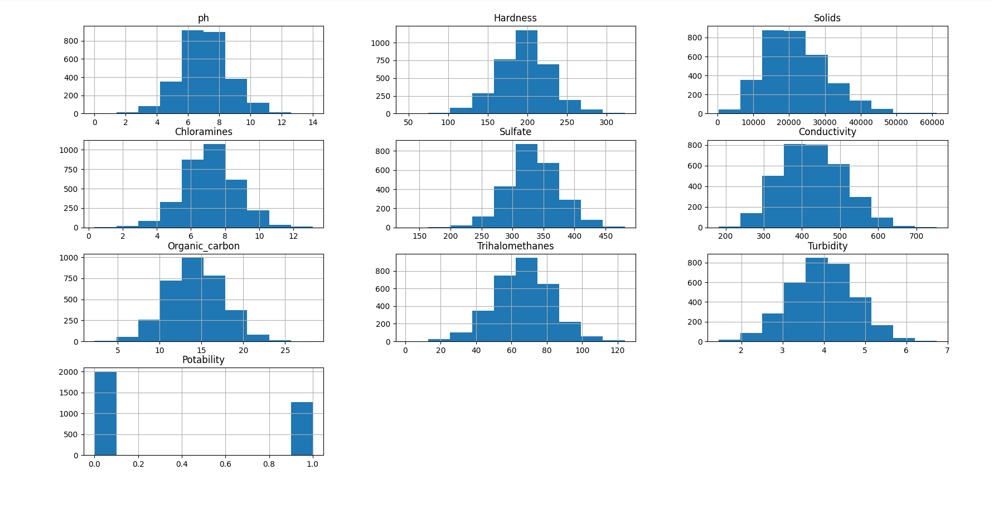
Countplot(Portability count)



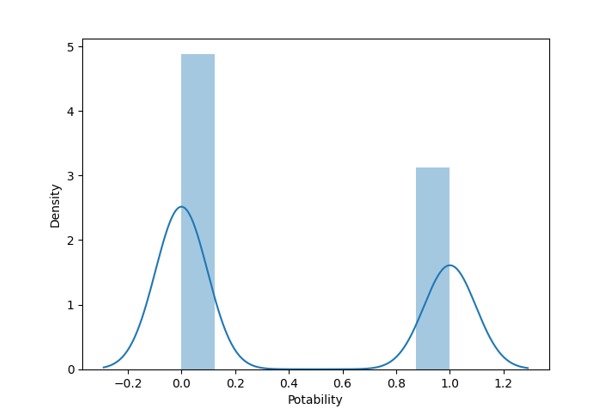
Violin plot(Portability vs pH)



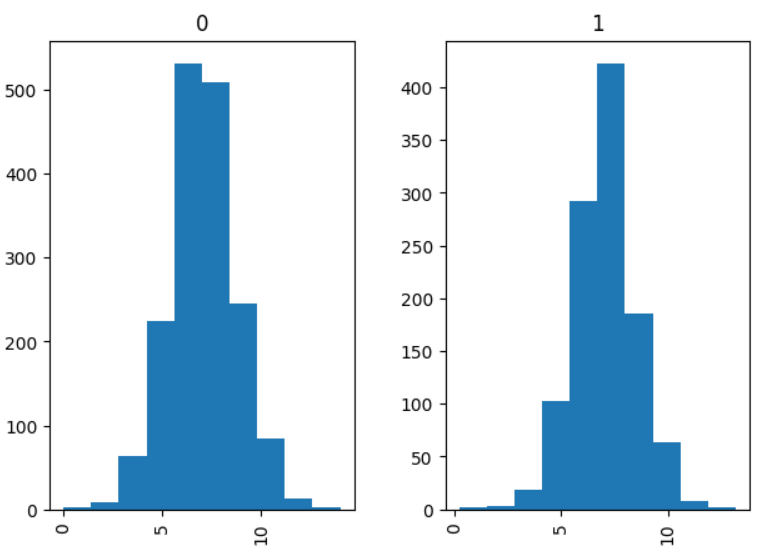
Sub plots



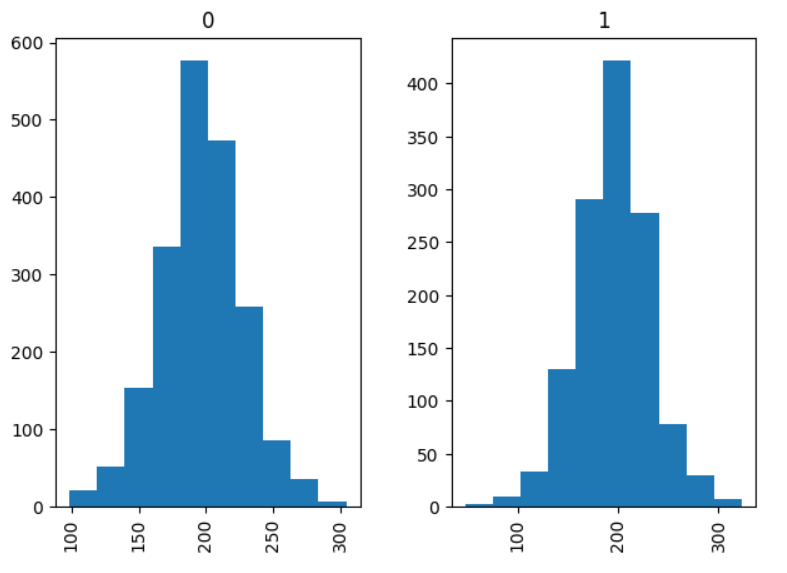
Histogram



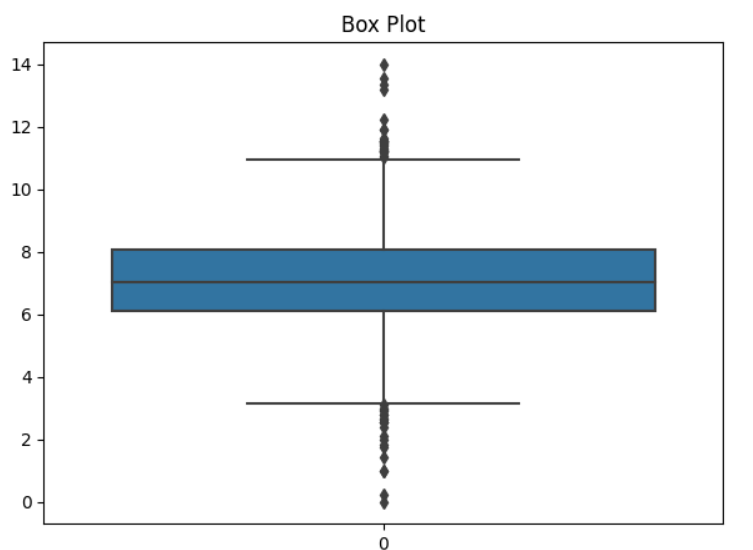
Portability vs Density



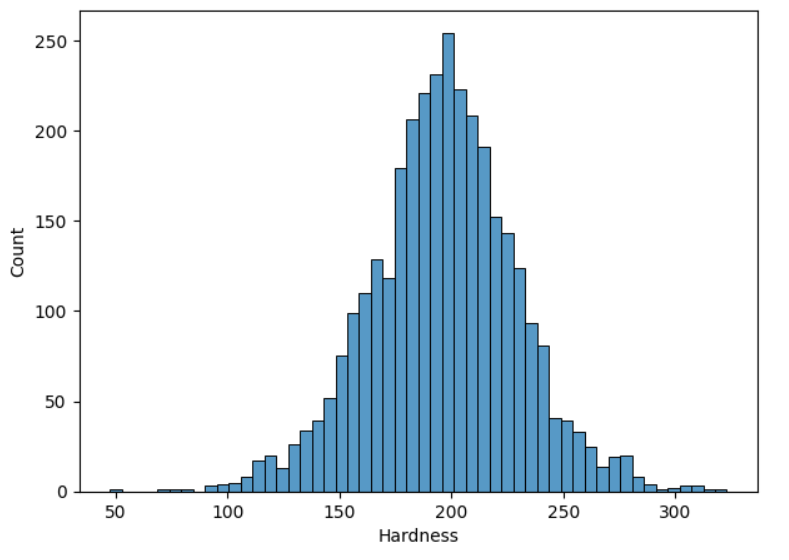
Portability vs pH



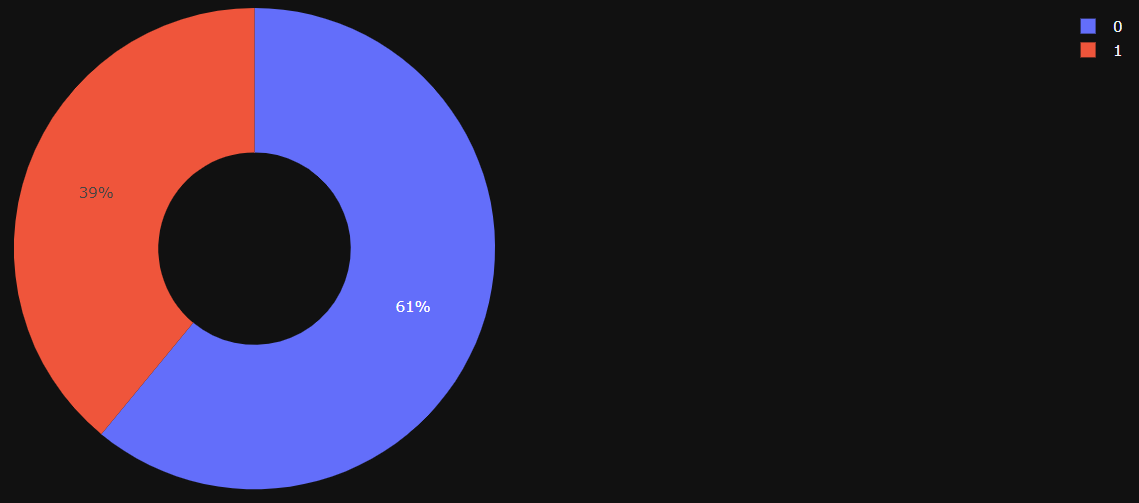
Portability vs Hardness



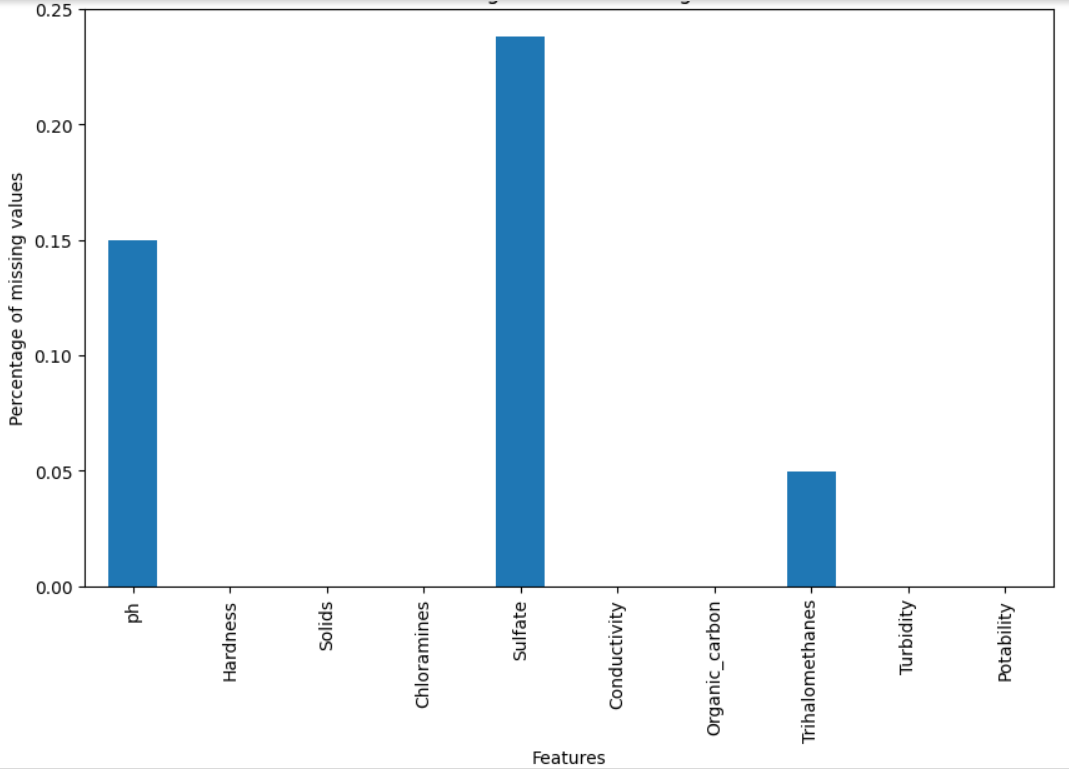
Boxplot



Hardness vs Count



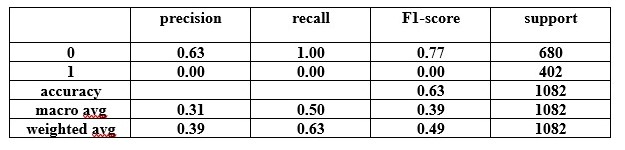
Portability pie chart



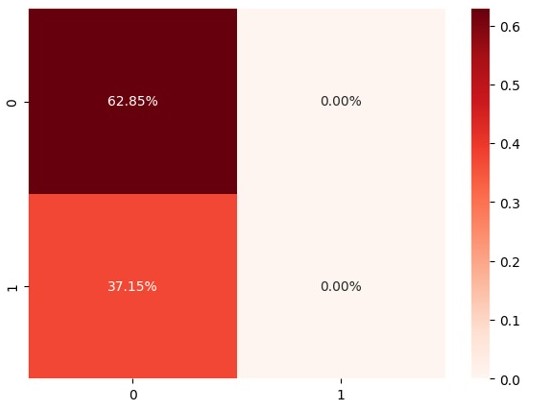
Missing data in percentages

6.1. LOGISTIC\_REGRESSION:

Logistic\_regression achieved accuracy score of 0.628466 . Logistic regression is linear model.

****

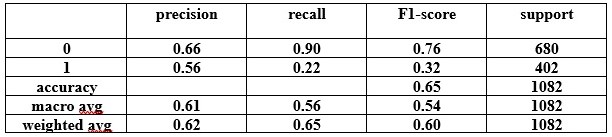
Report of logistic\_regression

****

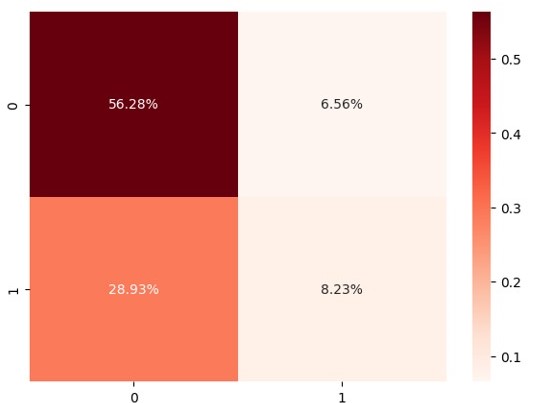
Confusion\_matrix

6.2.DECISION TREE:

DT algorithm secured accuracy of 0.645102 . Decision trees create hierarchical decision structures.

****

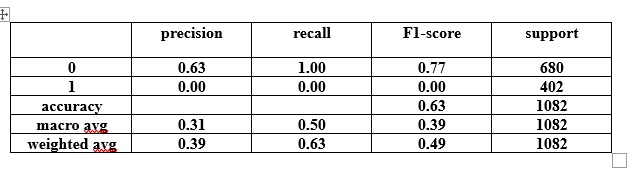
report of decision\_tree

****

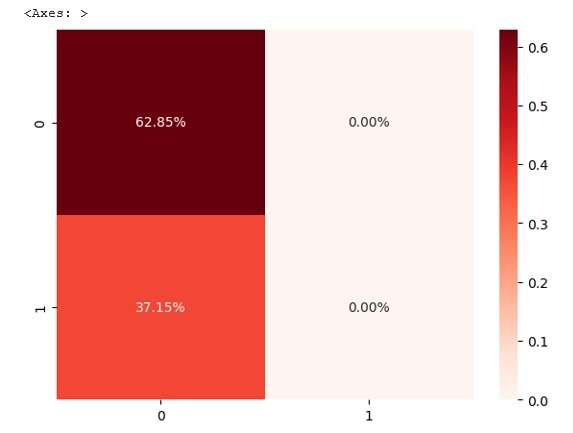
Confusion\_matrix

6.3 RANDOM FOREST :

RF and logistic\_regression shared same accuracy of 0.628466, indicating that random forest , despite being an ensemble method, did not significantly outperform logistic regression in this analysis.



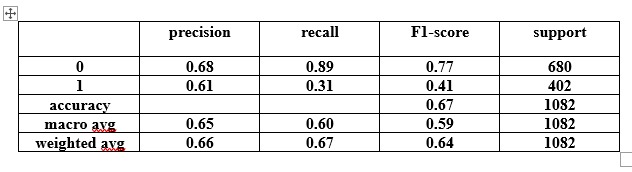
report of random\_forest



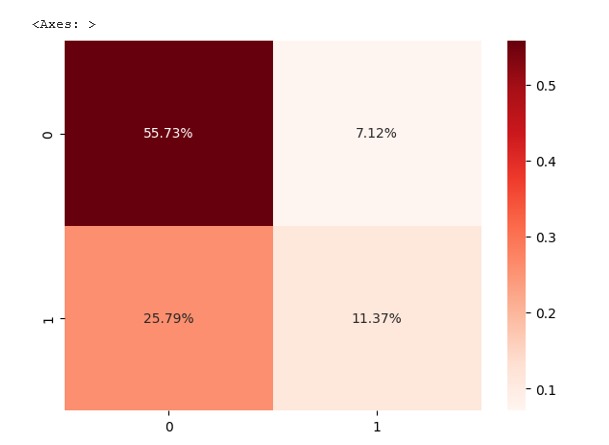
Confusion matrix

6.4 XG BOOST:

XG boost , a gradient boosting algorithm followed closely with an accuracy score of 0.670980.XG boost is a robust ensemble method that excels in predictive tasks and has proven to be a valuable choice in this analysis.



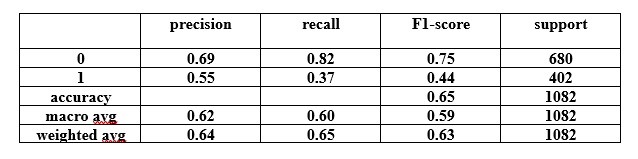
Report of XGBoost



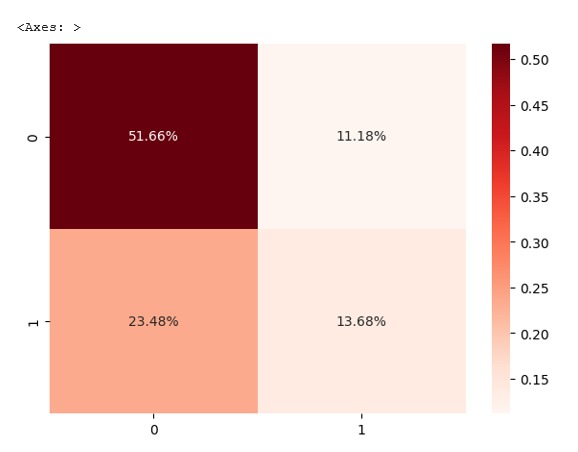
Confusion\_matrix

6.5 K NEAREST NEIGHBOUR:

K nearest neighbours achieved accuracy score of 0.653420 . KNN relies on local data patterns.



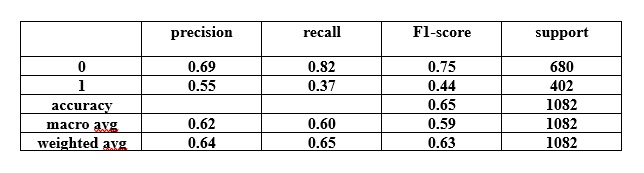
Report of knn



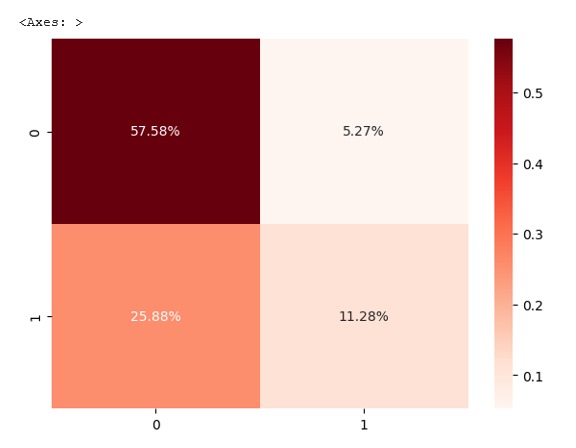
Confusion\_matrix

6.6 SVM:

SVM gained the more accuracy from algorithms with the score of 0.688540 .SVM is capable for its effectiveness in solving complex decision boundaries ,which may have contributed to its strong performance in this context.



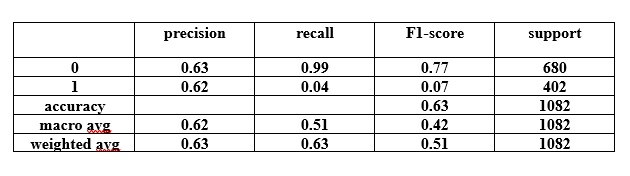
Report of svm



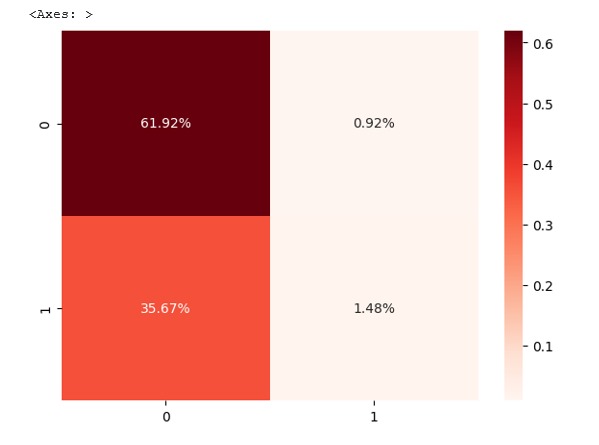
Confusion\_matrix

6.7 ADA BOOST:

Ada boost had accuracy score of 0.634011 .Ada boost is an ensemble learning method.



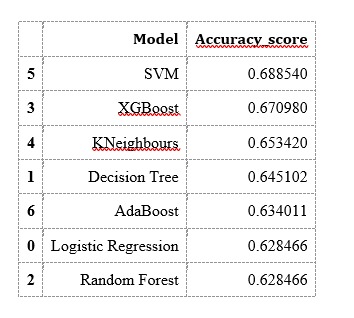
Report of ADABoost



Confusion\_matrix

6.8 MODEL ACCURACY SCORES:

The analysis of ML algorithms for water quality classification , along with their respective accuracy scores.

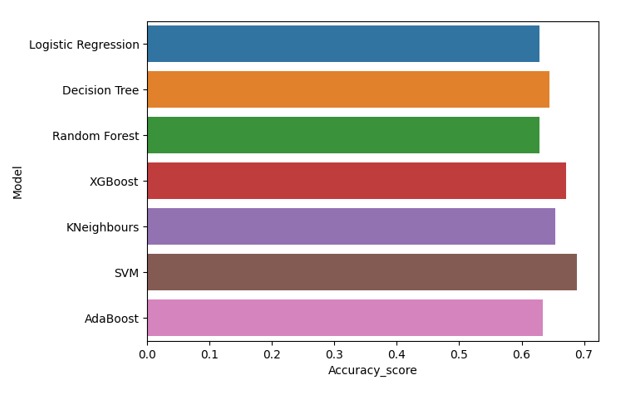


Accuracy scores

**7.CONCLUSION AND FUTURE SCOPE**

To estimate quality of water from collected data , we have used 7 ML algorithms: Logistic -regression , Decision-Tree, Random-Forest , XG Boost , K-NN , SVM and AdaBoost, demonstrating their effectivity in accurately predicting quality of water .

While utilising above applications of ML algorithms in water quality classification , showcased its potential of revolutionizing environmental monitoring and management.



The performance of these models , mostly relies on quality, representativeness of the data taken for train and test. If dataset contains errors, outliers, or biases ,it can significantly affect the models predictions.

The analysis does not explore the optimisation of hyper parameters for each model. Fine tuning hyper parameters can substantially improve model performance. The default settings used for some models may not be optimal.

The analysis does not address the issue of imbalanced classes , which is common in many real world classification problems. Techniques like resampling or using appropriate evaluation methods should be considered.

In conclusion, the SVM emerged as the top performer in predicting water quality based on provided data , closely followed by XG Boost .However, further analysis , including considerations of model interpretability , computational efficiency , and real world constraints, should be evaluated to process the most suitable algorithm, for deployment in a practical water quality prediction system.

**8.REFERENCES**

**[1]** Aliashrafi, A., Zhang, Y., Groenewegen, H., Peleato, N.M., 2021. A Review of Data-Driven Modelling in Drinking Water Treatment. Rev. Environ. Sci. Bio/Technol. 20 (4), 985–1009.

**[2]** Bedell, E., Harmon, O., Fankhauser, K., Shivers, Z., Thomas, E., 2022. A Continuous, inSitu, near-Time Fluorescence Sensor Coupled with a Machine Learning Model for Detection of Fecal Contamination Risk in Drinking Water: design, Characterization and Field Validation. Water Res., 118644

**[3]** Branch, A., Trinh, T., Carvajal, G., Leslie, G., Coleman, H.M., Stuetz, R.M., Drewes, J.E., Khan, S.J., Le-Clech, P., 2016. Hazardous Events in Membrane Bioreactors–Part 3: impacts on Microorganism Log Removal Efficiencies. J. Memb. Sci. 497, 514–523.

**[4]** Briscoe, E., Feldman, J., 2011. Conceptual Complexity and the Bias/Variance Tradeoff. Cognition 118 (1), 2–16.

**[5]** Cheswick, R., Cartmell, E., Lee, S., Upton, A., Weir, P., Moore, G., Nocker, A., Jefferson, B., Jarvis, P., 2019. Comparing Flow Cytometry with Culture-Based Methods for Microbial Monitoring and as a Diagnostic Tool for Assessing Drinking Water Treatment Processes. Environ. Int. 130, 104893.

**[6]** Diaz-Elsayed, N., Rezaei, N., Guo, T., Mohebbi, S., Zhang, Q., 2019. Wastewater-Based Resource Recovery Technologies across Scale: a Review. Resour. Conserv. Recycl. 145, 94–112.

**[7]** Dom`enech, L., Saurí, D., 2010. Socio-Technical Transitions in Water Scarcity Contexts: public Acceptance of Greywater Reuse Technologies in the Metropolitan Area of Barcelona. Resour. Conserv. Recycl. 55 (1), 53–62.

**[8]** Foschi, J., Turolla, A., Antonelli, M., 2021. Soft Sensor Predictor of E. Coli Concentration Based on Conventional Monitoring Parameters for Wastewater Disinfection Control. Water Res. 191, 116806.

**[9]** Garner, E., Inyang, M., Garvey, E., Parks, J., Glover, C., Grimaldi, A., Dickenson, E., Sutherland, J., Salveson, A., Edwards, M.A., 2019. Impact of Blending for Direct Potable Reuse on Premise Plumbing Microbial Ecology and Regrowth of Opportunistic Pathogens and Antibiotic Resistant Bacteria. Water Res. 151, 75–86.

**[10]** Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-Derived Soft-Sensors for Biological Wastewater Treatment Plants: an Overview. Environ. Model. Softw. 47, 88–107.

**[11]** Ramaraj, M., & Sivakumar, R. (2023). Integration of band regression empirical water quality (BREWQ) model with deep learning algorithm in spatiotemporal modeling and prediction of surface water quality parameters. Modeling Earth Systems and Environment, 1-26.

**[12]** Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. Water Research, 229, 119422.

**[13]** Han, M., Su, Z., & Na, X. (2023). Predict water quality using an improved deep learning method based on spatiotemporal feature correlated: a case study of the Tanghe Reservoir in China. Stochastic Environmental Research and Risk Assessment, 1-13.

**[14]** Ibrahim, H., Yaseen, Z. M., Scholz, M., Ali, M., Gad, M., Elsayed, S., ... & Khalifa, M. M. (2023). Evaluation and prediction of groundwater quality for irrigation using an integrated water quality indices, machine learning models and GIS approaches: A representative case study. Water, 15(4), 694.

**[15]** Yan, T., Zhou, A., & Shen, S. L. (2023). Prediction of long-term water quality using machine learning enhanced by Bayesian optimisation. Environmental Pollution, 318, 120870.

**[16]** Narita, K., Matsui, Y., Matsushita, T., & Shirasaki, N. (2023). Screening priority pesticides for drinking water quality regulation and monitoring by machine learning: Analysis of factors affecting detectability. Journal of Environmental Management, 326, 116738.

**[17]** Gaagai, A., Aouissi, H. A., Bencedira, S., Hinge, G., Athamena, A., Heddam, S., ... & Ibrahim, H. (2023). Application of water quality indices, machine learning approaches, and GIS to identify groundwater quality for irrigation purposes: a case study of Sahara Aquifer, Doucen Plain, Algeria. Water, 15(2), 289.

**[18]** Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2023). Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Safety and Environmental Protection, 169, 808-828.

**[19]** Zheng, H., Liu, Y., Wan, W., Zhao, J., & Xie, G. (2023). Large-scale prediction of stream water quality using an interpretable deep learning approach. Journal of environmental management, 331, 117309.

**[20]** Omeka, M. E., Igwe, O., Onwuka, O. S., Nwodo, O. M., Ugar, S. I., Undiandeye, P. A., & Anyanwu, I. E. (2023). Efficacy of GIS-based AHP and data-driven intelligent machine learning algorithms for irrigation water quality prediction in an agricultural-mine district within the Lower Benue Trough, Nigeria. Environmental Science and Pollution Research, 1-30.