WATER QUALITY PREDICTION USING MACHINE LEARNING ALGORITHMS

P. Krishna Prasad[1], Kishan Ranjit[2]

[1]Assistant Professor, Department of MCA, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India

[2]MCA Student, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India

Abstract-Assessment of river water quality (WQ) is one of the most important duties of authorities in worldwide water resource management. When establishing a water quality file (WQI), water evaluations take into account a number of quality-related criteria. In the age of sub-files, WQI evaluations are infamously tedious and prone to errors. The most recent machine learning (ML) approaches, known for their greater accuracy, may be used to address this problem. Water samples were taken from the wells in the region under consideration (North Pakistan) in order to create WQI expectation models. Four independent calculations were employed in this study: M5P, irregular trees (RT), random woods (RF), and diminishing mistake pruning tree (REPT). Twelve half-and-half information mining calculations, including independent sacking (BA), randomizable sifted grouping (RFC), and cross-approval border determination (CVPS), were also applied. The data was split into two groups (70:30) using the 10-crease cross-approval procedure before the outcomes were computed. Ten random input permutations with Pearson correlation coefficients were performed in order to find the best dataset combination for improving the algorithm's prediction. Low correlation variables performed poorly, despite the fact that hybrid algorithms were able to predict outcomes better than many independent algorithms.

KEYWORDS:machine learning, hybrid algorithms, Prediction

I. INTRODUCTION

Vital backslide is a genuine methodology that is used for building simulated intelligence models where the dependent variable is dichotomous: Water contamination is one of the fundamental difficulties of the cutting edge presence where the objectives like the Accumulated Countries Reasonable Improvement Objectives (UN-SDGs) and a wise and judicious planet are being sought after. Pasquale De Meo was the accomplice administrator responsible for sorting out the review of this synthesis and endorsing its appropriation. All social orders, ecologies, and creations are subject to cultivating. drinking, sterilization, and the development of energy The worldwide water emergency is one of the significant dangers to humankind at the present time. Subsequently, groundwater sum and quality are gigantic overall concerns. A few sicknesses, including cholera, free entrail disorder, typhoid, amebiasis, hepatitis, gastroenteritis, giardiasis, campylobacteriosis, scabies, and worm illnesses, are welcomed on by defiled water. Detachment of the entrails was the justification behind practically 1.6 million passings in 2017 alone. Harms in water significantly affect the climate, which thusly influences the soundness of people and marine life.

The unloading of present day waste, pesticides, and composts, as well as uncontrolled and misguided urbanization, all add to water pollution. This sort of defilement is more clear in streams or streams that are near new advancements in metropolitan regions. With both non-ceaseless point sources, stream pollution is transforming into a genuinely serious issue that examines experts in overall water the board. This sort of tainting really brings down the nature of

1.

the water (WQ). Sea life and the accessibility of clean water for drinking and agrarian use are fundamentally affected by WQ defilement. In non-modern nations, which a significant part of the time experience monetary ups and downs, it is all the more determinedly to settle the tainting issue. Moreover, every improvement action could have essential ordinary repercussions. For instance, the necessity for seriously cultivating creation descends on the normal wealth of soils as a result of an extension in people and interest for extra resources. Accordingly, engineered manures become more important to support yield. Feces that isn't needed is routinely dumped into endlessly streams, dirtying ground and underground water sources. Hence, there is a creating interest for WQ assessment and checking. WQ noticing and evaluation are essential for the protection of human prosperity, the climate, and the native environment. Fortunate, valuable, and long-range water the bosses' arrangements can achieve this. The water quality file (WQI) is utilized to assess the WQ. WQI helps guide policymakers' exercises and decisions. Regardless, because of the responsibility of various sub-records and conditions, deciding WQI is definitely not a fundamental correspondence. WQI is a record that isn't layered and is made by obvious WQ factors. Various elements that are used are pH (capacity of hydrogen), DO (separated oxygen), TSS (hard and fast suspended solids), Body (natural oxygen interest), AN (ammoniacal nitrogen), and COD (substance oxygen interest). WQ may be surveyed without a second thought due to the assessment associations. Assessing factors like Ca2+, Mg2+, NO3, and others are essential for normal assessments of groundwater quality markers (GQIs). The evaluation of WQ incorporates two or three pieces of water, including physical, material, run of the mill, and radiological. WQI is moreover a customarily elaborate way for concluding whether WQ the heap up measures are feasible or insufficient. WQIs incorporate the English Columbia Water Quality Record (BCWQI), the Oregon Water Quality File (OWQI), the Florida Stream Water Quality List (FWQI), the In-between time Public Water Quality Norms for Malaysia (INWQS), the Canadian Water Quality File (CQI), and the US Public Disinfection Establishment Water Quality File (NSFWQI). WQI is determined through various techniques and computations worldwide.ch as double Information and the connection between a solitary ward variable and at least one free factors can be portrayed utilizing strategic relapse. The independent variables can be apparent, ordinal, or of stretch sort.

The possibility of the calculated capability that it utilizes is the wellspring of the expression "strategic relapse." The essential ability is generally called the sigmoid capacity. The value of this essential ability lies some place in the scope of nothing and one.

II. LITERATURE SURVEY

Xu dong Jia et al. [1] have used both descriptive analysis and machine learning to examine the quality of the water. They start by obtaining the data source from the Kaggle website. After data processing, we perform data mining using the Python sklearn module. The first step is the selection of the machine learning data mining technique using description analysis. The decision tree, Bayesian algorithm, and KNN are the methods we ultimately use to analyze the water data from the Kaggle website. By using a machine learning technique, the data will be divided into available and unavailable categories. Finally, using these three approaches, they have obtained the outcomes of three approaches and conduct a matching comparison and analysis.

K Abirami et al. [2] worked on Water Quality Index (WQI), which serves as a single number to identify the quality of water, would be used in the proposed study to evaluate the water quality. A unique class called the Water Quality Class (WQC) will be created based on the WQI result. pH, temperature, conductivity, dissolved oxygen (DO), biological oxygen demand (BOD), nitrate, and total coliform are the variables used to calculate the water quality index (WQI). While there are numerous machine learning algorithms available for categorization, it is essential to pick the best one. In order to compare the performance of the K-Nearest Neighbor (K-NN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, and Random Forest algorithms, various evaluation measures, including Accuracy score, Confusion Matrix, Precision, Recall, and f1-score, were used.

Bilal Aslam [3] in this study, water samples were taken from wells in the study area (Northern Pakistan) to develop WQI prediction models. Four independent algorithms, i.e., random trees (RT), random forest (RF), M5P, and reduced error pruning (REPT), were used in this study. In addition, 12 hybrid data mining algorithms (combination of discrete, bagging (BA), cross-validation parameter selection (CVPS) and randomized filtered classification (RFC)) were used. Using a 10-fold cross-validation technique, the data were divided into two groups (70:30) to generate the algorithm. Ten random input permutations were generated using Pearson's correlation coefficients to identify the best possible combination of data sets to improve the algorithm's prediction. Variables with very low correlations performed poorly, while hybrid algorithms improved the predictive power of multiple independent algorithms.

Vinoth Kumar P et al. [4] has studied that indicates potential improvements after analyzing previous water quality prediction studies. In this study, a new intelligent aquaculture system using a machine intelligence model (SAS-MI) was analyzed and proposed in previous water quality prediction works. The new SAS-MI model uses a deep convolutional neural network (D-CNN) and a k-means clustering technique. The k-means clustering used in SAS-MI aggregates the unlabeled data set used for training and testing. D-CNN predicts water quality for intelligent aquaculture using neural automatic feature technology. The performance of the proposed model was evaluated through a comparative study conducted with existing prediction models such as logistic regression, decision trees, XG boost classifiers, k-nearest neighbors and SVM. The SAS-MI model with D-CNN and k-means clustering provides significant results in terms of prediction accuracy, F1 score, and mean squared error (MSE).

Priyanshu Rawat et al. [5] studied provides a comprehensive analysis of the effectiveness of eight different machine learning algorithms in predicting water quality. Algorithms including Gaussian Naive Bayes, Extreme Gradient Boost classifier, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest and Decision Tree were tested using the potable water dataset. The main goal of this study was to find the best accurate machine learning algorithm for water quality prediction and to provide a comprehensive comparison of these methods. Algorithmic efficiency. The results of the study showed that one algorithm performed better than the others, with the lowest root mean square error and the highest accuracy.

Jaswanth Reddy Vilupuru et al. [6] in paper uses artificial intelligence techniques to predict water quality index (WQI) and water quality classification (WQC). Water quality data from India was used in this article. Neural network models such as long short term memory (LSTM) and regression models such as Ridge Regression, Random Forest Regressor with Randomized search CV have been developed to predict WQI. Machine learning models like KNN, Logistic Regression, Logistic Regression using GridSearchCV, XGBoost, SVM and SVM using Grid SearchCV for train test distributions like 70-30, 80-20 were used in WQC predictions. WQI prediction results showed that Ridge regression achieved the best R^2 of 95.21% with an MSE of 0.11. In WQC predictions, XGBoost achieved the highest accuracy (97.48%).

Wildan Azka Fillah et al. [7] studied on a benchmark water quality model using ARIMA, SVR and LSTM. It showed that LSTM algorithm gave the best result with less error. The model of the LSTM method can be used to make predictions, such as a seven-day forecast of the next day's pH value, whether it follows the rules or whether it needs to be checked. The company is not penalized for this preventive maintenance.

Sheng Cao et al. [8] has focused on water quality pollution, builds a water quality assessment model to analyze the water quality level, and provides an objective additional forecast on the development of its factors. In that paper, the genetic algorithm mutation factor is incorporated into the PSO algorithm. A least squares support vector machine (LS-SVM) based on an adaptive particle swarm optimization (PSO) algorithm for hyperparameter optimization creates a single water quality classification evaluation model. The fuzzy data granulation method is combined with LS-SVR (Least Square Support Regression) to create a water quality time series model that can predict the changing trend of water quality data over three days. Thanks to theoretical analysis and experimental data, this estimation model and prediction algorithm is faster in terms of training speed and accuracy compared to the traditional BP neural network.

M Uma Maheswari et al. [9] has investigated various supervised machine learning algorithms for water quality detection. Several variables are important in determining water quality, such as pH, hardness, solids, chloramines, sulfates, conductivity, organic carbon, trihalomethanes, turbidity, and potability. water Random Forest (RF) and Decision Tree (DT) are used to determine the caliber of water suitable for human consumption. The standard laboratory method of testing water quality is time consuming and can sometimes be expensive. The algorithms proposed in this study are able to provide an assessment of drinking water quality in a very short period of time. DT height accuracy F1 is 99% while RF score is 87.86% and accuracy is 82.36%. The difference between these two scores is because the accuracy of DT is lower. The proposed method shows its potential for use in real-time water quality monitoring systems, achieving adequate accuracy with a minimal set of parameters. This is necessary to demonstrate the usefulness of this program.

Suma S et al. [10] has developed predictive model to identify water samples that require further analysis to make the lab technician's work more efficient. WEKA software was used to implement the model, based on secondary data collected from the Kenya Water Institute. Water samples were classified into clean and polluted categories using a decision tree algorithm. When evaluating water quality, the determining factor is its alkalinity and conductivity. Public health and safety depend on the availability of clean drinking water. The researchers used five decision tree classifiers to evaluate the accuracy of the model: J48, LMT, Random Forest, Hoeffding Tree and Decision Stump.

EXISTING SYSTEM

It was found that the SVR model produced the best outcomes utilizing two estimations: a backslide tree (RT) computation and an assist vector with backsliding (SVR) estimation. Kayaalp et al. developed a SVR model based on crossbreeds. to gauge WQI using month-to-month WQ boundary information and the firefly estimation (FFA). The computation exhibited a significant expansion in expectation execution when contrasted with the independent SVR model. By diminishing the SVM calculation, Kamyab-Talesh et al. looked into the most important factors that affect the WQI. The creators guarantee that nitrate is the main element for WQI expectation. Wang and co. investigated three ML computations: SVR, SVR-GA (genetic calculation), and SVR-PSO (atom swarm streamlining) to estimate WQI and assess their presentation. Choice tree-based techniques, as M5P, RF, RT, REPT, and others, need stowed away units and can create displaying results that are better than those of ANFIS and ANN.

Disadvantages:-

Not well in prediction accuracy.
Its will not supported with dynamic changes.

PROPOSED METHOD

After the first data collection, several WQ metrics may be extracted from the water samples. The data were then added to datasets for verification and testing. The optimum info mix was identified from the testing datasets. In the conclusion, several algorithms were applied to forecast WQI on the best kinds, and the optimum algorithm was determined after an evaluation of the algorithms.

The most notable expectation power is found in the computations RF, credulous bayeis, strategic relapse (LR), and decision tree (DT). All computations were accepted as the estimated WQI for each model for each testing dataset was compared to the anticipated WQI.

Advantages:-

Improved effectiveness and exactness

ML calculations is a strong information demonstrating device that can catch and address complex info/yield connections.

Ready to scope with huge dataset.

Precise expectation.

III. SYSTEM ARCHITECTURE

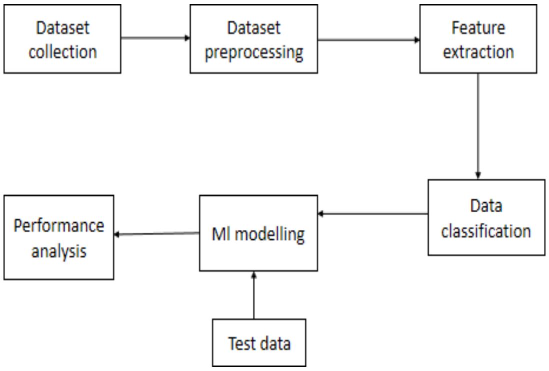


Figure 1

IV. METHODOLOGY

i. Data Gathering,
ii. preprocessing of the data,
iii. feature extraction,
iv. evaluation model, and
v. user interface

Data Gathering

This paper's information assortment comprises of various records. The determination of the subset of all open information that you will be working with is the focal point of this stage. Preferably, ML challenges start with a lot of information (models or perceptions) for which you definitely know the ideal arrangement. Marked information will be data for which you are as of now mindful of the ideal result.

Pre-Processing of Data

Format, clean, and sample from your chosen data to organise it.

There are three typical steps in data pre-processing:

1. *Designing*
2. *Information cleaning*
3. *Inspecting*

Designing: It's conceivable that the information you've picked isn't in a structure that you can use to work with it. The information might be in an exclusive record configuration and you would like it in a social data set or text document, or the information might be in a social data set and you would like it in a level document.

Information cleaning; is the most common way of eliminating or supplanting missing information. There can be information examples that are inadequate and come up short on data you assume you really want to resolve the issue. These events could should be eliminated.Moreover, a portion of the traits might contain delicate data, and it very well might be important to antonymize or totally eliminate these properties from the information.

Inspecting: You might approach significantly more painstakingly picked information than you want. Calculations might take significantly longer to perform on greater measures of information, and their computational and memory prerequisites may likewise increment. Prior to considering the whole datasets, you can take a more modest delegate test of the picked information that might be fundamentally quicker for investigating and creating thoughts.

Feature Extraction

The following stage is to A course of quality decrease is include extraction. Highlight extraction really modifies the traits instead of element choice, which positions the ongoing ascribes as indicated by their prescient pertinence. The first ascribes are straightly joined to create the changed traits, or elements. Finally, the Classifier calculation is utilized to prepare our models. We utilize the Python Normal Language Tool stash's classify module.

We utilize the gained marked dataset. The models will be surveyed utilizing the excess marked information we have. Pre-handled information was ordered utilizing a couple of AI strategies. Irregular woodland classifiers were chosen. These calculations are generally utilized in positions including text grouping.

Assessment Model

Model The method involved with fostering a model incorporates assessment. Finding the model that best portrays our information and predicts how well the model will act in what's to come is useful. In information science, it isn't adequate to assess model execution utilizing the preparation information since this can rapidly prompt excessively hopeful and overfitted models. Wait and Cross-Approval are two procedures utilized in information science to evaluate models.

The two methodologies utilize a test set (concealed by the model) to survey model execution to forestall over fitting. In light of its normal, every classification model's presentation is assessed. The result will take on the structure that was envisioned. diagram portrayal of information that has been ordered.

ALGORITHMS:

1) Logistic regression

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.Logistic regression accuracy is 66%.

2) Support vector machine:

Support Vector Machines, or SVMs for short, are classification and regression machine learning algorithms. SVMs are one of the strong AI calculations for arrangement, relapse and anomaly recognition purposes. A model is built by an SVM classifier, and new data points are assigned to one of the categories that are given. In this manner, it tends to be seen as a non-probabilistic double straight classifier.

Linear classification is a possible application for SVMs. Using the kernel trick, SVMs can effectively perform non-linear classification in addition to linear classification. It empower us to certainly plan the contributions to high layered include spaces.

Hyperplane:

A hyperplane is a choice limit what isolates between given set of information focuses having different class marks. Using the widest possible hyperplane, the SVM classifier divides the data points. This hyperplane is known as the most extreme edge hyperplane and the straight classifier it characterizes is known as the greatest edge classifier.

Support Vectors:

Support vectors are the example data of interest, which are nearest to the hyperplane. By calculating margins, these data points will better define the separating line or hyperplane.

Margin The distance that separates the two lines on the closest data points is called a margin. It is determined as the opposite separation from the line to help vectors or nearest pieces of information. In SVMs, we strive to achieve maximum margin by maximizing this separation gap.

SVC is the another implementation of the SVM.

Svm accuracy 77%.

3) Naive Bayes

The Naive Bayes Computation is one of the basic estimations in simulated intelligence that helps with request issues. It is gotten from Bayes' probability speculation and is used for text portrayal, where you train high-layered datasets. The Naive Bayes Algorithm is useful for a variety of tasks, including spam filtering, sentiment analysis, and article classification.

Request estimations are used for arranging novel discernment into predefined classes for the unenlightened data. The Innocent Bayes Calculation is well-known for its simplicity and sufficiency. With this calculation, models can be constructed and expectations can be made quicker.

Navie bayes algorithm accuracy 66%.

4) KNN

KNN is one of the most straightforward AI calculations given the Managed Learning methodology. The new case is placed in an arrangement that is largely comparable to the accessible orders by KNN computation, which anticipates the similarity between the new case/data and open cases.The KNN computation maintains all relevant data and groups additional relevant data based on proximity. This shows that new data will typically be simply grouped into a well-suited class using KNN computation when it first appears. Similar to how it is utilized for characterisation difficulties, the KNN computation can be applied for Relapse and Order. KNN is a non-parametric calculation, therefore it doesn't assume anything about the data below. Additionally, it is known as a dormant student computation since it keeps the dataset rather than quickly acquiring it from the readiness set and then doing a computation on it when requested.

KNN computation simply stores the dataset during the ready stage and groups new data into a grouping that is comparable to the dataset when new data is received.

KNN accuracy is 74%.

5) Decision Tree

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It is a tree-like model where each internal node represents a decision based on a feature (attribute), each branch represents an outcome of that decision, and each leaf node represents the final prediction or outcome. The tree structure allows the algorithm to make a sequence of decisions to arrive at a final prediction for a given input.Decision trees work on these 2 phases,'Training' Phase and 'Prediction' Phase.

Decision trees are versatile and widely used in various domains due to their simplicity and ability to handle both classification and regression tasks. However, to overcome some of their limitations, ensemble methods like Random Forests or Gradient Boosting are often employed, which combine multiple decision trees to improve accuracy and robustness.Decision tree accuracy is 99%.

6) Random Forest

An AI technique called Random Forest is outfit-based and operated. You can combine various computation types to create a more convincing forecast model, or use a similar learning technique at least a few times. The phrase "Irregular Timberland" refers to how the arbitrary woodland method combines a few calculations of the same type or different chosen trees into a forest of trees. The irregular timberland technique can be used for both relapse and characterization tasks..

Coming up next are the essential stages expected to execute the irregular woods calculation.

Pick N records aimlessly from the datasets.

Utilize these N records to make a choice tree.Select the number of trees you that need to remember for your calculation, then, at that point, rehash stages 1 and 2.

Each tree in the timberland predicts the classification to which the new record has a place in the order issue. The classification that gets most of the votes is at last given the new record.The Advantages of Irregular Woodland.The way that there are numerous trees and they are completely prepared utilizing various subsets of information guarantees that the irregular timberland strategy isn't one-sided.The irregular woods strategy fundamentally relies upon the strength of "the group," which reduces the framework's general predisposition. Since it is extremely challenging for new information to influence every one of the trees, regardless of whether another information point is added to the datasets, the general calculation isn't highly different. In circumstances when there are both downright and mathematical highlights, the irregular woods approach performs well.At the point when information needs esteems or has not been scaled, the irregular woodland method likewise performs well. In this project random forest accuracy is 89%.

7) AdaBoost

Versatile helping is a method utilized for parallel grouping. We use short decision trees as weak learners to implement AdaBoost.

AdaBoost implementation steps:

1. Train the base model utilizing the weighted preparation information

2. The next step is to add weak learners in order to make it a strong learner.

3. A decision tree is the component of each weak learner; dissect the result of every choice tree and allocate higher loads to the misclassified results. With higher weights, this gives the prediction more weight.

4. Proceed with the interaction until the model becomes equipped for anticipating the exact outcome

Adaboost algorithm accuracy is 98%.

8) XGBoost

XGBoost is a superior scattered slant helping library expected for successful and adaptable planning of computer based intelligence models. A gathering learning method creates a more grounded forecast by joining the expectations of numerous feeble models. One of the most popular and widely used machine learning algorithms is XGBoost, which stands for "Extreme Gradient Boosting." It can handle large datasets and perform at the cutting edge in many classification and regression tasks.One of the most important aspects of XGBoost is its capacity to deal with real-world data with missing values without requiring a significant amount of pre-processing. Likewise, XGBoost comes furnished with worked in help for equal handling, making it conceivable to prepare models on monstrous datasets in a sensible measure of time. Among the many applications of XGBoost are click-through rate prediction, recommendation systems, and Kaggle competitions. It is furthermore incredibly versatile and considers changing of various model limits to further develop execution.

Xgboost algorithm accuracy is 99%

User Interface:

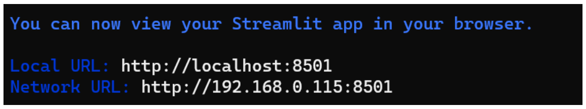
The pattern of Information Science and Examination is expanding step by step. From the information science pipeline, one of the main advances is model sending. We have a ton of choices in python for sending our model. A few well known systems are Carafe and Django. Yet, the issue with utilizing these systems is that we ought to have some information on HTML, CSS, and JavaScript. Remembering these requirements, Adrien Treuille, Thiago Teixeira, and Amanda Kelly made "Streamlit". Presently utilizing streamlit you can send any AI model and any python project easily and without stressing over the frontend. Streamlit is very easy to use.

In this article, we will get familiar with a few significant elements of streamlit, make a python project, and convey the task on a nearby web server. How about we introduce streamlit. Type the accompanying order in the order brief.

pip install streamlit

When Streamlit is introduced effectively, run the given python code and in the event that you don't get a mistake, then streamlit is effectively introduced and you can now work with streamlit. Instructions to Run Streamlit record:

How to Run Streamlit file:



```
You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.0.115:8501
```

Figure 2

V. CONCLUSION FUTURE SCOPE

The five well-known data mining classification methods utilized in this work to categorize water quality as good, acceptable, somewhat polluted, polluted, and seriously polluted are Naive Bayes, Decision tree, K-nearest neighbor, Support Vector

Machines, and Random Forest. Each classifier's models were built on top of the overall index of pollution.The synthetic data set was made using the eight possible ranges of the following parameters: temperature, conductivity, dissolved oxygen (DO), pH, biochemical oxygen demand (BOD), nitrates (NO3), feces, and total coli (TC) forms. These ranges complied with both national and international norms. The real data set was derived from the literature that was available for a number of Tamil Nadu areas.

Every classifier's boundaries were calibrated during the learning phase to appear at the ideal boundary settings for discovering a particular water quality class in the informative indices. Metrics like accuracy, sensitivity, specificity, recall, and F1score were employed in the testing phase to judge each predictive model's efficacy against hypothetical data. Out of the five alternatives available, the Radom Forest classifier yields the best results. The DT classifier also discovered the RF performance level.

The Statistical and ML algorithms were used in this research that provided highly accurate results; it will be beneficial to use deep learning algorithms, for instance, convolution neural network, to cross-check the results and compare them with this study to yield holistic results.

Results:

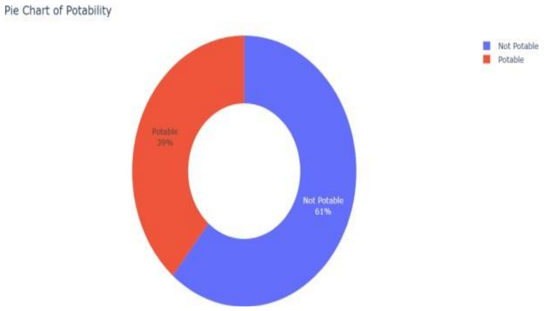


Figure 3

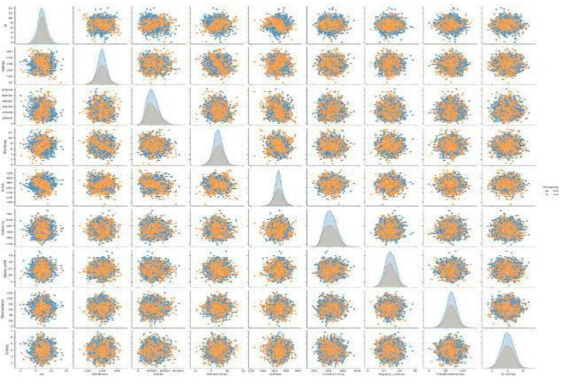


Figure 4

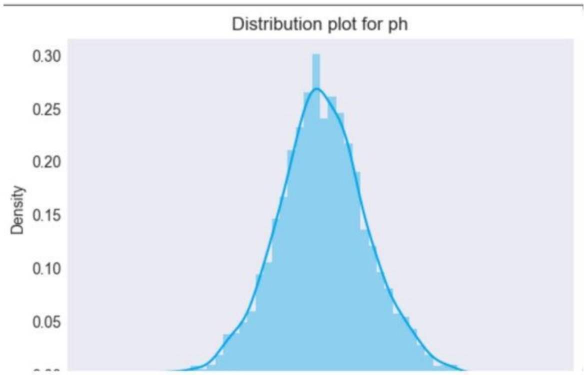


Figure 5

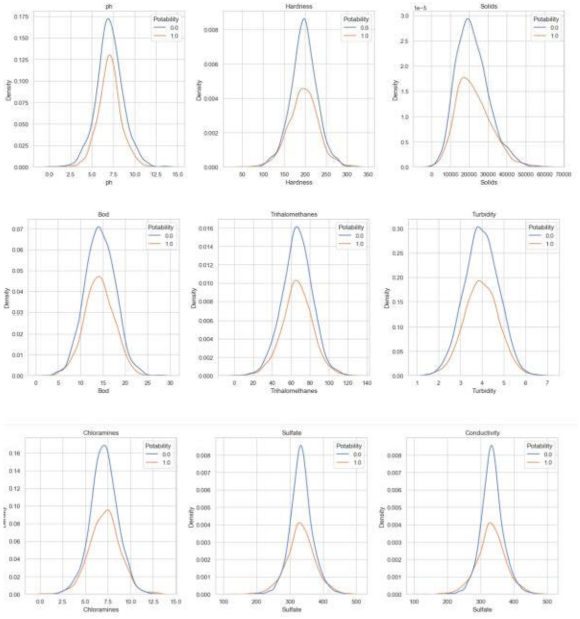
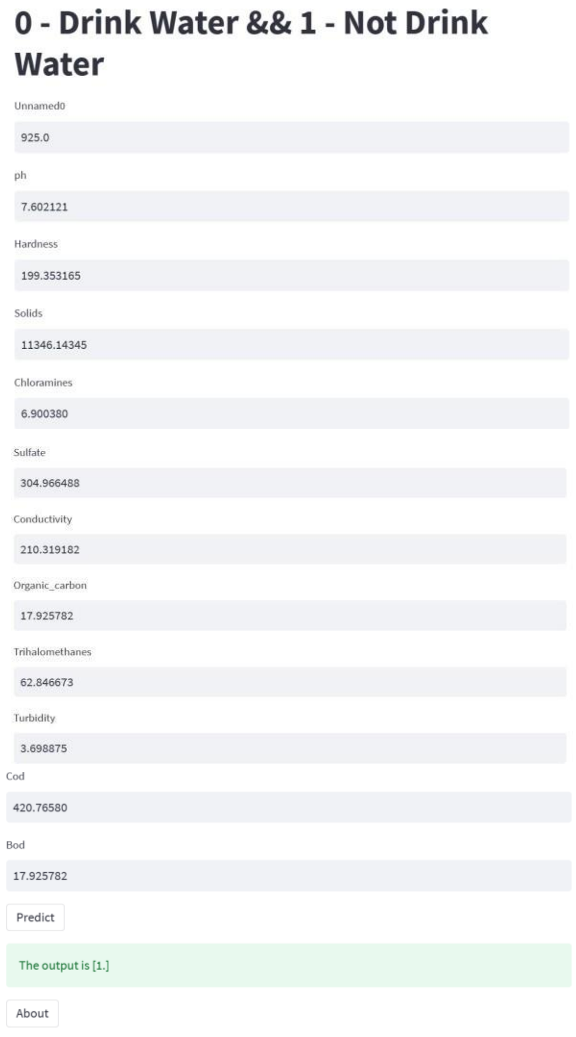


Figure 6

Water quality prediction

Streamlit water quality prediction using ML App

Figure 7



0 - Drink Water && 1 - Not Drink Water

Unnamed0
925.0

ph
7.602121

Hardness
199.353165

Solids
11346.14345

Chloramines
6.900380

Sulfate
304.966488

Conductivity
210.319182

Organic_carbon
17.925782

Trihalomethanes
62.846673

Turbidity
3.698875

Cod
420.76580

Bod
17.925782

The output is [1.]

Figure 8

REFERENCES:

[1] X. Jia, "Detecting Water Quality Using KNN, Bayesian and Decision Tree," 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), Hangzhou, China, 2022, pp. 323-327, doi: 10.1109/CACML55074.2022.00061.

[2] K. Abirami, P. C. Radhakrishna and M. A. Venkatesan, "Water Quality Analysis and Prediction using Machine Learning," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 241-245, doi: 10.1109/CSNT57126.2023.10134661.

[3] B. Aslam, A. Maqsoom, A. H. Cheema, F. Ullah, A. Alharbi and M. Imran, "Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach," in IEEE Access, vol. 10, pp. 119692-119705, 2022, doi: 10.1109/ACCESS.2022.3221430.

[4] V. K. P, S. K, B. M. D and R. Reshma, "Predicting and Analyzing Water Quality using Machine Learning for Smart Aquaculture," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 354-359, doi: 10.1109/ICSCDS56580.2023.10104677.

[5] P. Rawat, M. Bajaj, V. Sharma and S. Vats, "A Comprehensive Analysis of the Effectiveness of Machine Learning Algorithms for Predicting Water Quality," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 1108-1114, doi: 10.1109/ICIDCA56705.2023.10099968.

[6] J. R. Vilupuru, D. C. Amuluru and G. B. K, "Water Quality Analysis using Artificial Intelligence Algorithms," 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2022, pp. 1193-1199, doi: 10.1109/ICIRCA54612.2022.9985650.

[7] W. A. Fillah and D. Purwitasari, "Prediction of Water Quality Index using Deep Learning in Mining Company," 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2022, pp. 1-5, doi: 10.1109/ICITISEE57756.2022.10057870.

[8] S. Cao, S. Wang and Y. Zhang, "Design of River Water Quality Assessment and Prediction Algorithm," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018, pp. 901-906, doi: 10.1109/ICMLA.2018.00146.

[9] M. U. Maheswari, R. Sudharsanan, M. Arthy, A. Jenefer, L. Oormila and V. Samuthira Pandi,

"Efficient Drinking Water Quality Analysis using Machine Learning Model with Hyper-Parameter Tuning," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 401-406, doi: 10.1109/ICICCS56967.2023.10142799.

[10] M. U. Maheswari, R. Sudharsanan, M. Arthy, A. Jenefer, L. Oormila and V. Samuthira Pandi, "Efficient Drinking Water Quality Analysis using Machine Learning Model with Hyper-Parameter Tuning," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 401-406, doi: 10.1109/ICICCS56967.2023.10142799.

[11] M. Awais, B. Aslam, A. Maqsoom, U. Khalil, F. Ullah, S. Azam, andM. Imran, ‘‘Assessing nitrate contamination risks in groundwater: A machine learning approach,’’ Appl. Sci., vol. 11, no. 21, p. 10034,Oct. 2021.

[12] T. H. Tulchinsky and E. A. Varavikova, ‘‘Communicable diseases,’’ New Public Health, San Diego, CA, USA, Tech. Rep., 2014, p. 149.

[13] S. Khalid, M. Shahid, I. Bibi, T. Sarwar, A. H. Shah, and N. K. Niazi,‘‘A review of environmental contamination and health risk assessment of wastewater use for crop irrigation with a focus on low and high-income countries,’’ Int. J. Environ. Res. Public Health, vol. 15, no. 5, p. 895, May 2018.

[14] E. Chu and J. Karr, ‘‘Environmental impact: Concept, consequences, mea-surement,’’ Reference Module Life Sci., Elsevier, 2017.

[15] P. M. Kopittke, N. W. Menzies, P. Wang, B. A. McKenna, and E. Lombi, ‘‘Soil and the intensification of agriculture for global food security,’’Environ. Int., vol. 132, Nov. 2019, Art. no. 105078.

[16] M. Hameed, S. S. Sharqi, Z. M. Yaseen, H. A. Afan, A. Hussain, and A. Elshafie, ‘‘Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia,’’ Neural Comput. Appl., vol. 28, pp. 893–905, Dec. 2017.

[17] T. Bournaris, J. Papathanasiou, B. Manos, N. Kazakis, and K. Voudouris, ‘‘Support of irrigation water use and eco-friendly decision process in agricultural production planning,’’ Oper. Res., vol. 15, no. 2, pp. 289–306,Jul. 2015.

[18] F. Rufino, G. Busico, E. Cuoco, T. H. Darrah, and D. Tedesco, ‘‘Evaluating the suitability of urban groundwater resources for drinking water and irrigation purposes: An integrated approach in the Agro-Aversano area ofSouthern Italy,’’ Environ. Monitor. Assessment, vol. 191, no. 12, pp. 1–17, Dec. 2019.

[19] M. Vadiati, A. Asghari-Moghaddam, M. Nakhaei, J. Adamowski, andA. H. Akbarzadeh, ‘‘A fuzzy-logic based decision-making approach for identification of groundwater quality based on groundwater quality indices,’’ J. Environ. Manage., vol. 184, pp. 255–270, Dec. 2016.

[20] A. Shalby, M. Elshemy, and B. A. Zeidan, ‘‘Assessment of climate change impacts on water quality parameters of lake Burullus, Egypt,’’ Environ. Sci.Pollut. Res., vol. 27, no. 26, pp. 32157–32178, Sep. 2020.

[21] M. Awais, B. Aslam, A. Maqsoom, U. Khalil, F. Ullah, S. Azam, and M. Imran, ‘‘Assessing nitrate contamination risks in groundwater: A machine learning approach,’’ Appl. Sci., vol. 11, no. 21, p. 10034, Oct. 2021.

[22] T. H. Tulchinsky and E. A. Varavikova, ‘‘Communicable diseases,’’ New Public Health, San Diego, CA, USA, Tech. Rep., 2014, p. 149.

[23] S. Khalid, M. Shahid, I. Bibi, T. Sarwar, A. H. Shah, and N. K. Niazi, ‘‘A review of environmental contamination and health risk assessment of wastewater use for crop irrigation with a focus on low and high-income countries,’’ Int. J. Environ. Res. Public Health, vol. 15, no. 5, p. 895, May 2018.

[24] E. Chu and J. Karr, ‘‘Environmental impact: Concept, consequences, mea- surement,’’ Reference Module Life Sci., Elsevier, 2017.

[25] P. M. Kopittke, N. W. Menzies, P. Wang, B. A. McKenna, and E. Lombi, ‘‘Soil and the intensification of agriculture for global food security,’’ Environ. Int., vol. 132, Nov. 2019, Art. no. 105078.

[26] M. Hameed, S. S. Sharqi, Z. M. Yaseen, H. A. Afan, A. Hussain, and A. Elshafie, ‘‘Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia,’’ Neural Comput. Appl., vol. 28, pp. 893–905, Dec. 2017.

[27] T. Bournaris, J. Papathanasiou, B. Manos, N. Kazakis, and K. Voudouris, ‘‘Support of irrigation water use and eco-friendly decision process in agricultural production planning,’’ Oper. Res., vol. 15, no. 2, pp. 289–306, Jul. 2015.

[28] F. Rufino, G. Busico, E. Cuoco, T. H. Darrah, and D. Tedesco, ‘‘Evaluating the suitability of urban groundwater resources for drinking water and irrigation purposes: An integrated approach in the Agro-Aversano area of Southern Italy,’’ Environ. Monitor. Assessment, vol. 191, no. 12, pp. 1–17,

Dec. 2019.

[29] M. Vadiati, A. Asghari-Moghaddam, M. Nakhaei, J. Adamowski, and A. H. Akbarzadeh, ‘‘A fuzzy-logic based decision-making approach for identification of groundwater quality based on groundwater quality indices,’’ J. Environ. Manage., vol. 184, pp. 255–270, Dec. 2016.

[30] A. Shalby, M. Elshemy, and B. A. Zeidan, ‘‘Assessment of climate change impacts on water quality parameters of lake Burullus, Egypt,’’ Environ. Sci.Pollut. Res., vol. 27, no. 26, pp. 32157–32178, Sep. 2020.

[31] D. Sharma and A. Kansal, ‘‘Water quality analysis of river Yamuna using water quality index in the national capital territory, India (2000–2009),’’ Appl. Water Sci., vol. 1, nos. 3–4, pp. 147–157, Dec. 2011.

[32] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, ‘‘Improving prediction of water quality indices usingnovel hybrid machine-learning algorithms,’’ Sci. Total Environ., vol. 721,Jun. 2020, Art. no. 137612.

[33] Z. M. Yaseen, M. M. Ramal, L. Diop, O. Jaafar, V. Demir, and O. Kisi, ‘‘Hybrid adaptive neuro-fuzzy models for water quality index estimation,’’ Water Resour. Manage., vol. 32, pp. 2227–2245, May 2018.

[34] C. Iticescu, L. P. Georgescu, G. Murariu, C. Topa, M. Timofti, V. Pintilie, and M. Arseni, ‘‘Lower Danube water quality quantified through WQI and multivariate analysis,’’ Water, vol. 11, no. 6, p. 1305, Jun. 2019.

[35] W. C. Leong, A. Bahadori, J. Zhang, and Z. Ahmad, ‘‘Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM),’’ Int. J. River Basin Manage., vol. 19,no. 2, pp. 149–156, Apr. 2021.

[36] I. H. Sarker, ‘‘Machine learning: Algorithms, real-world applications and research directions,’’ Social Netw. Comput. Sci., vol. 2, no. 3, pp. 1–21, May 2021.

[37] M. J. Alizadeh, M. R. Kavianpour, M. Danesh, J. Adolf, S. Shamshirband, and K.-W. Chau, ‘‘Effect of river flow on the quality of estuarine and coastal waters using machine learning models,’’ Eng. Appl. Comput. Fluid Mech., vol. 12, no. 1, pp. 810–823, Jan. 2018.

[38] K. Kargar, S. Samadianfard, J. Parsa, N. Nabipour, S. Shamshirband, A. Mosavi, and K.-W. Chau, ‘‘Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms,’’ Eng. Appl. Comput. Fluid Mech., vol. 14, no. 1, pp. 311–322, Jan. 2020.

[39] T. M. Tung and Z. M. Yaseen, ‘‘A survey on river water quality modelling using artificial intelligence models: 2000–2020,’’ J. Hydrol., vol. 585, Jun. 2020, Art. no. 124670.

[40] K. Khosravi, L. Mao, O. Kisi, Z. M. Yaseen, and S. Shahid, ‘‘Quantifying hourly suspended sediment load using data mining models: Case study of a glacierized Andean catchment in Chile,’’ J. Hydrol., vol. 567, pp. 165–179, Dec. 2018.