Water Quality Assessment Through Predictive Machine Learning

1st Md.Shahriar Hossain Apu
Dept.of IoT and Robotics Engineering
Bangabandhu Sheikh Mujibur Rahman Digital University
Bangaldesh
1901036@iot.bdu.ac.bd

2nd Chowdhury Rafsan Mohammadullah
Dept.of IoT and Robotics Engineering
Bangabandhu Sheikh Mujibur Rahman Digital University
Bangaldesh
1801048@iot.bdu.ac.bd

***Abstract*—Freshwater is a critical resource for agriculture and industry's survival. Examination of water quality is a fundamental stage in the administration of freshwater assets. As indicated by the World Health Organization's yearly report, many individuals are getting sick or some are dead due to the lack of safe drinking water, especially pregnant ladies and kids. It is critical to test the quality of water prior to involving it for any reason, whether it is for animal watering, chemical spraying(Pesticides etc), or drinking water. Water quality testing is a strategy for finding clean drinking water. Accordingly, appropriate water monitoring is basic for safe, clean, and sterile water. Water testing is fundamental for looking at the legitimate working of water sources, testing the safety of drinking water, identifying disease outbreaks, and approving methodology and safeguard activities. Water quality is a proportion of a water's readiness for a specific utilize in view of physical, chemical, and biological qualities.**

***Index Terms*—Water Quality prediction, SVM, Random Forest, Decision Tree, Logistic Regression, Hyperparameter Tuning, KNN, Naive Bayes.**

I. Introduction

Water is the principal source for shipping energy to each cell in the body and is additionally the regulator of all body capacities. The cerebrum contains 80% of water. Extreme drying out may prompt mental hindrances and loss of capacity to obviously think. Water is one of the most fundamental regular assets for the endurance of the whole life on this planet. In light of the nature of water, it tends to be utilized for various purposes like drinking, washing, or water system. Plants and creatures likewise rely upon water for their endurance. To put it plainly, all living organic entities need an enormous amount and great nature of water for presence. Freshwater is a fundamental asset to horticulture and industry for its essential presence. Water quality observation is a key stage in the administration of freshwater assets. As indicated by the yearly report of WHO, many individuals are kicking the bucket because of the absence of unadulterated drinking water particularly pregnant ladies and youngsters. It is critical to check the nature of water for its expected reason, whether it be animals watering, compound showering, or drinking water. Water quality testing is a device that can be utilized to find unadulterated drinking water. Consequently, the right checking of water is incredibly much significant for protecting unadulterated, and clean water. Water testing assumes a key part in breaking down the right activity of water supplies, testing the wellbeing of drinking water, perceiving sickness flare-ups, and approving cycles and precaution measures. Water quality is the proportion of the reasonableness of water for a specific reason in view of explicit physical, substance, and organic attributes. Testing the nature of a water body, both surface water, and groundwater, can assist us with responding to inquiries concerning whether the water is satisfactory for drinking, washing, or water system to give some examples of applications. It can utilize the consequences of water quality tests to look at the nature of water starting with one water body and then onto the next in a locale, state, or across the entire country. Microbiological quality is for the most part the main pressing concern on the grounds that irresistible infections brought about by pathogenic microorganisms, infections, helminths, and so on are the most well-known and boundless wellbeing risk connected with drinking water. Overabundance amount of certain synthetic substances in drinking water prompts well-being risk. These synthetics incorporate fluoride, arsenic, and nitrate. Safe drinking (consumable) water should be passed on to the client for drinking, food game plan, individual neatness, and washing. The water ought to satisfy the normal quality rules for making it pure at the spot of supply to the clients. One of the greatest machine (statistical) learning algorithms for pragmatic applications is Breiman's random forests (RF). Despite its practical usefulness, random forests remained fairly obscure until recently when compared to other AI and machine learning techniques, with little use in water research, particularly hydrological applications. As a result, the power of 'Breiman's' unique calculation and its variations in water assets, and applications remains unutilized. Aside from the standard applications of RF-based calculations in relapse and grouping issues, as well as the calculation of significant measurements, heir utilization for decile expectation, endurance investigation, and offhand surmising, appear to be less well known among water researchers and professionals. Random forest is accepted to have a place with the class of data driven models with regards to water resources.Literature Survey shows a few current studies on the use of information-driven models in water assets, water asset design, and hydrology, where the random forest is lacking and a significant portion of the text is dedicated to brain

organizations. Some of the most often discussed topics in literature on information-driven models are expectation (or determination), preprocessing, variable selection, splitting the dataset into preparation and testing phases, and predictive execution assessments

II. LITERATURE SURVEY

[9]In 2018, Ali Heidar Nasrolahi along with Amir Hamzeh Haghiabi and Abbas Parsaie predicted the Water Quality of a river bed in Iran Tireh River by taking pH, Na, Ca, Mg, etc such components into consideration. Performance was tallied by using several ML and DL algorithms. It was observed that results of SVM was the front runner and gave the best accuracy. ANN gave acceptable accuracy for practical purposes.

[6]In 2019, Umair Ahmed et.al explained ways to efficiently predict water quality using supervised Machine Learning. Harrowing diseases have been in increased proportions due to the depreciation and deterioration of water quality at an alarming rate which was a direct impact of rapid urbanization and industrialisation. Their research monitors and works with supervised Machine Learning algorithms to calculate Water Quality Index (WQI) and Water Quality Class (WQC), the former being a singular index which describes the general quality of water and the latter being the derivative and distinctive class on the basis of WQI.

[7]In 2020, Mohammed Al-Yaari et.al illustrated the use of Artificial Intelligence algorithms along with the performance of each used algorithm. As we know, for the protection of the environment, predicting and modelling of the quality of water is immensely important. In the methodology they proposed, to predict WQI, artificial intelligence algorithms, such as, NARNET and LSTM were used. Along with this, KNN, SVM and Naïve Bayes algorithms were also implemented. They used a dataset with 7 relevant and significant features and statistical parameters were used to develop the model and evaluate them.

[12]In 2020, Navideh Noori et.al explained the water quality prediction using SWAT-ANN coupled approach. For solving environmental problems Machine Learning algorithm such as Artificial Neural Networks is being used widely. They illustrated the application of SWAT-ANN for water quality prediction.

[13]In 2022, Jin-Won Yu et.al explained the use of AI algorithms for the water quality prediction. Combined the power of data decomposition, fuzzy C-means clustering and bidirectional gated recurrent model for the prediction of water quality.

In 2022, [11] Manisha Koranga et.al discussed the use of Machine Learning Algorithms for water quality prediction for Nanital Lake, Uttarakhand. Analysed the use of machine learning algorithms and used eight regression algorithms and nine classification algorithms. Three algorithms Random Forest, SVM and Stochastic Gradient Descent comes out to be the most effective machine learning algorithms.

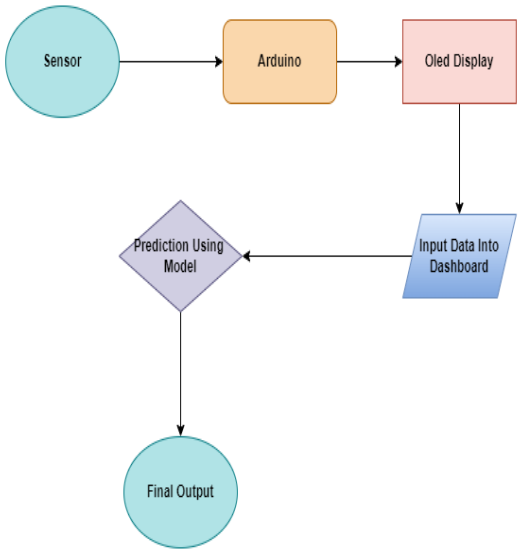

Fig. 1. System Architecture

In this paper [8] focuses on using machine learning models to predict and classify water quality, distinguishing between potable and non-potable water. The research uses a dataset of 3277 samples collected over 9 years from locations in Andhra Pradesh, India. Various models, including G-Naive Bayes, B-Naive Bayes, SVM, KNN, X Gradient Boosting, and Random Forest, were tested. The Random Forest model achieved the highest accuracy at 78.96%, while SVM had the lowest at 68.29%. This research highlights the potential of machine learning for effective water quality assessment and emphasizes the importance of selecting the best model for ensuring safe drinking water.

Reviewing the literature shows that artificial intelligence techniques have been proposed for water conservation projects for which water quality prediction and assessment plays an important role. Hence, this paper presents a designed algorithm for the prediction of Water Quality considering the concentration, pH, duration factors.

III. METHODOLOGY

To improve the water quality, pre-processing phase plays a vital role in data analysis. For the calculation of the Water Quality Index, the most significant factors are taken into consideration. For the system's superior accuracy, Data Normalization Techniques has been used.

A. DATA COLLECTION

The dataset used in this study for water quality assessment and classification was obtained from [10] Kaggle, a popular online platform for sharing and discovering datasets.By utilizing this dataset from Kaggle, the study gains access to a

rich source of real-world data, allowing for the evaluation of various machine learning models to predict and classify water quality effectively. The large and diverse nature of the dataset ensures that the findings of this research are robust and applicable to a wide range of water quality scenarios.

B. HARDWARE SETUP

For our system we used several sensor like turbidity,pH, water-level sensor,Analog Oxygen sensor,TDS sensor and as a micro-controller we used Arduino uno.

IV. System Learning

We trained our using some well known Machine learning algorithm. These are-

A. SUPPORT VECTOR MACHINE

[1]One of the most popular supervised learning algorithms is Support Vector Machine, it can be used for Regression and classification problems. Widley, it is used for Classification problems in Machine Learning. Creation of the decision boundary (which is the best area or plane or line) that helps to sort n-dimensional data space into classes. This helps us to put the new query point in the accurate category in the future. Whenever there's a new query point, it is compared to the decision boundary and is classified accordingly. This is the main goal of Support Vector Machine. Decision boundary which will suit the best for a particular dataset is called a Hyperplane. Hyperplane consists of extreme points/vector. Extreme cases indicate datapoints which lie in all the extremities such data points are termed as support vectors. Since, the whole algorithm is based on these extremities it referenced as Support Vector Machine. In the SVM algorithm, we use the loss function it helps us to the maximize the margin hinge loss.

B. DECISION TREE

[3]Decision tree is a supervised machine learning algorithm which can be used for both classification and regression. It is mostly preferred to solve classification problems in which data has to be classified into different categories. It has a tree like structure with internal node acting as features, branches as decision rule and leaf nodes as the outcome. It is basically a graphical representation of all the possible outcomes for a given problem based on the given conditions. It simply asks a question and based on the answers it further splits the trees. In order to build a tree, it uses CART algorithm.

C. RANDOM FOREST

[5]Random Forest is a classifier that contains number of decision tree on various subsets. It is a supervised machine learning algorithm used for both classification and regression. To improve the predicted accuracy of the dataset it takes the average of the decision trees. It is based on the concept of ensemble learning, which is basically a process of combining several classifiers to solve complex problems and to improve the performance of the model. To have the higher accuracy and prevent the problem of overfitting, greater number of trees are used.

D. KNN

[2]K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

E. Naïve Bayes Classifier

[4]Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

V. DATA PROCESS FLOW

There are basically 10 steps for making our model predict the water quality of the water samples. Those steps are:-

A. Problem Identification

In this step, we identify the problem which is solved by our model. So the problem to be solved by our model is water quality prediction using a dataset.

B. Data Extraction

In this, we extract the data from the internet to train our data and predict the water quality. So for that, we take the CPCB(Central Pollution Control Board India) dataset which contains 3277 instances of 13 different wellsprings which are collected between 2014 to 2020.

C. Data Exploration

In this step, we analyze the data visually by comparing some parameters of water with the WHO standards of water. It gives a slight overview of the data.

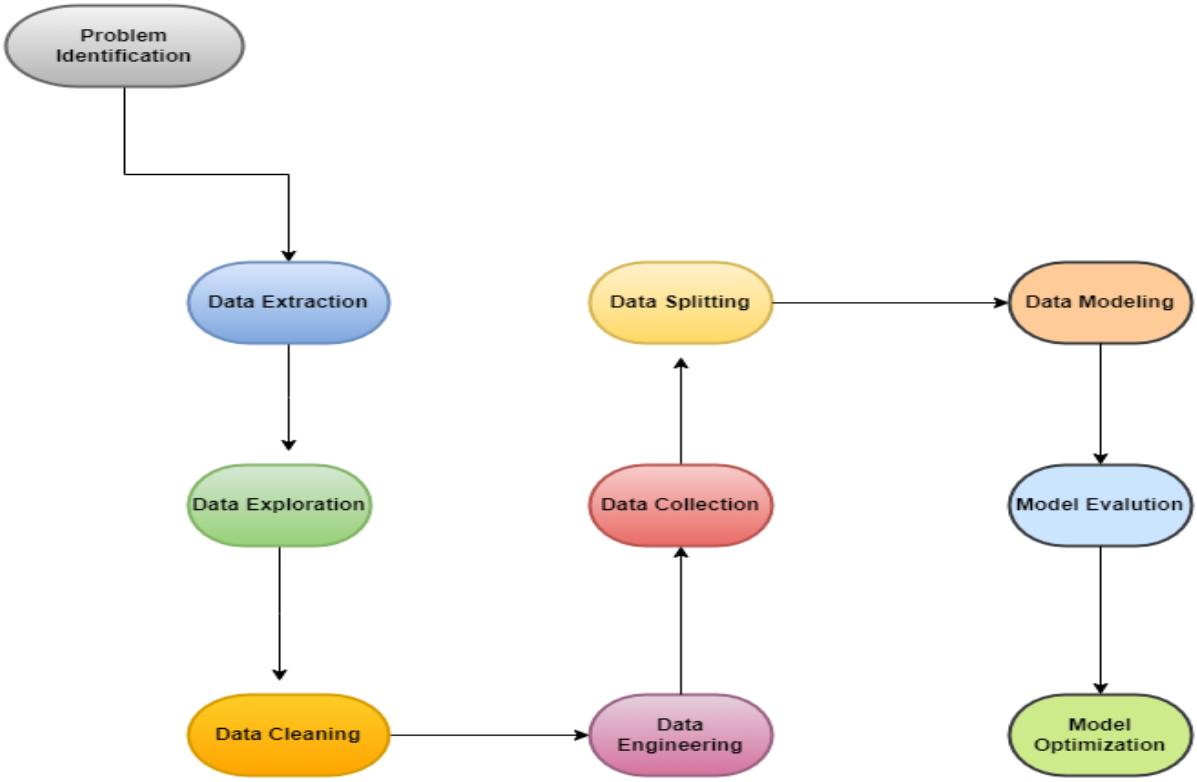

Fig. 2. Data Flow Diagram

D. Data Cleaning

In this step, we clean that data like if there are some missing values in it so we replace them with mean and remove noise from the data.

E. Data Engineering

In this step, we ensure that the data is quality data so that the prediction accuracy increases.

F. Data Selection

In this step, we select the data types and source of the data. The essential goal of data selection is deciding fitting data type, source, and instrument that permit agents to respond to explore questions sufficiently

G. Data Splitting

In this step, we divide the dataset into smaller subsets for easing the complexity. Normally, with a two-section split, one section is utilized to assess or test the information and the other to prepare the model.

H. Data Modeling

In this step, we create a graph of the dataset for visual representation of data for better understanding. A Data Model is this theoretical model that permits the further structure of conceptual models and to set connections between data.

I. Model Evaluation

Model Evaluation is a fundamental piece of the model improvement process. In this step, we evaluate our model and check how well our model do in the future.

VI. DATASET PARAMETER

[10]Parameters are the basic and most important component of a model. Based on the parameter model prediction and it shows the skill of the model over the data. Similarly, random forest algorithms also have some parameters to predict water quality. We take 10 parameter to predict the water quality. Those parameters are :-

1) Ph
2) Solid
3) Cholride
4) Conductance
5) Sulphate
6) Hardness as CaCO3
7) Trihalomethanes
8) Turbidity
9) Organic Carbon

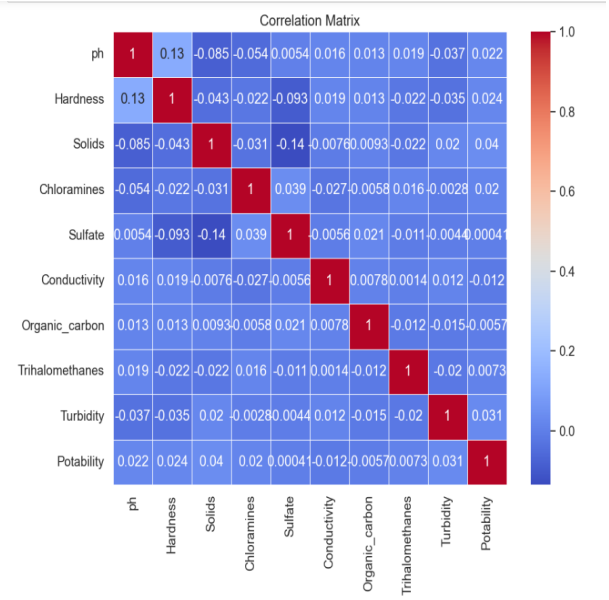

Fig. 3. Co relation Matrix of attribute

10) Potability

To predict whether the water is drinkable or not WHO provides some standard values for these parameters of water which is as follows,

TABLE I
WHO STANDARD FOR PARAMETERS

Parameter	WHO limit
Ph	6
Solid	500ppm
Cholride	200mg/l
Conductance	2000 macro S/cm Fecal Col
Sulphate	500mg/l
Hardness as CaCO3	500mg/l
Trihalomethanes	0.5ppb
Turbidity	1NTU
Organic Carbon	2mg/l
Potability	1 macrog/l

VII. DATA VISUALIZATION

In this step several approach is taken to understand the nature of data.Figure-3 shows the Co-relation of the attribute.

VIII. WATER QUALITY INDEX

WQI is the correlation of the sum with an erratic or logical norm or with a pre-determined base. In this way, the WQI observed and announced natural status and patterns on guidelines quantitatively. A water quality list is a way to sum up a lot of water quality information into straightforward terms (e.g., great) for answering to the board and the general population in a predictable way. Notwithstanding the nonattendance of a universally acknowledged composite file of water quality, a few nations have utilized and are involving collected water

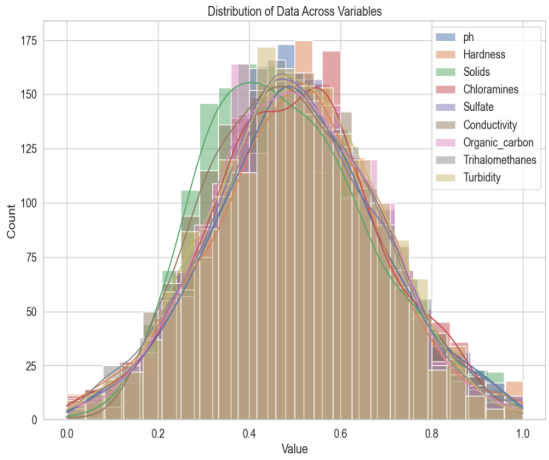

Fig. 4. Data Distribution Plot

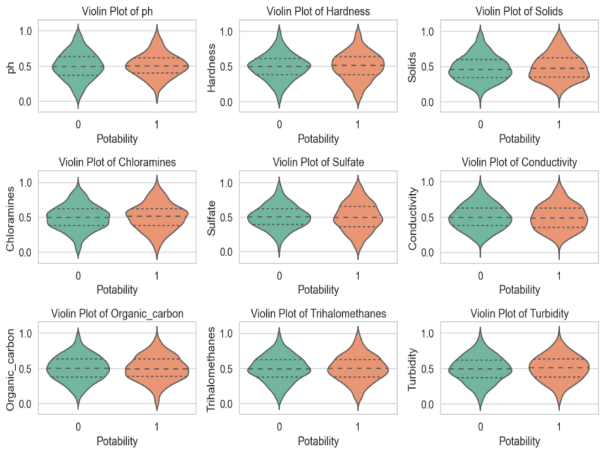

Fig. 5. Violin Plot

quality information in the improvement of water quality lists. To calculate the water quality index(WQI) conventionally we take 10 features of water to reflect the quality of water like ph, chloride, conductance, etc. In this paper, we use all 10 parameters to calculate the WQI of the water. The general formula to calculate the water quality index is given below,

where q is the boundary of that parameter, w-factor is the weight of that parameter. On the basis of WQI value it is determined that the quality of the water is drinkable or not.

IX. CONFUSION MATRIX

For classification problems confusion matrix is used on a wide scale. it is used for multiclass classification problems and

$$WQI = \frac{\sum q_{value} \times w\_factor}{\sum w\_factor}$$

Fig. 6. Equation of water Quality Index

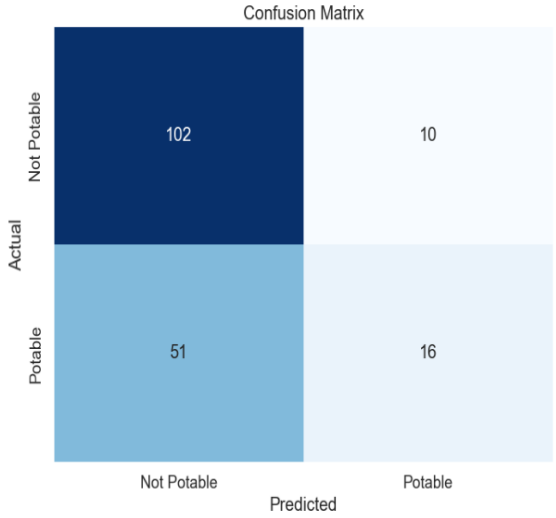

Fig. 7. Confusion Matrix

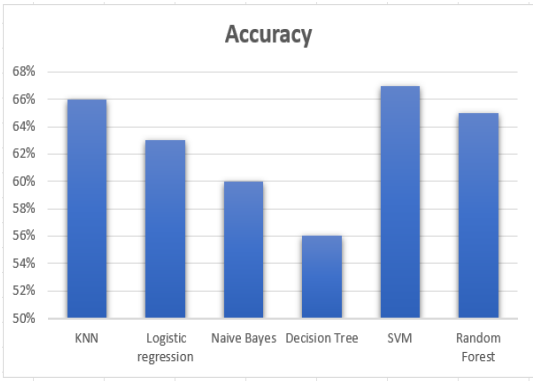

Fig. 8. Accuracy of the Model

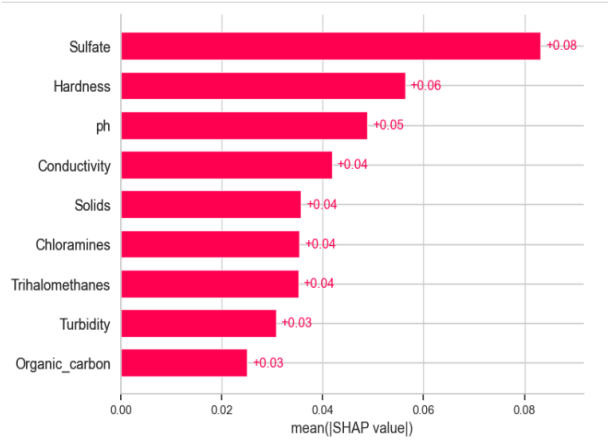

Fig. 9. SHAP Bar Graph

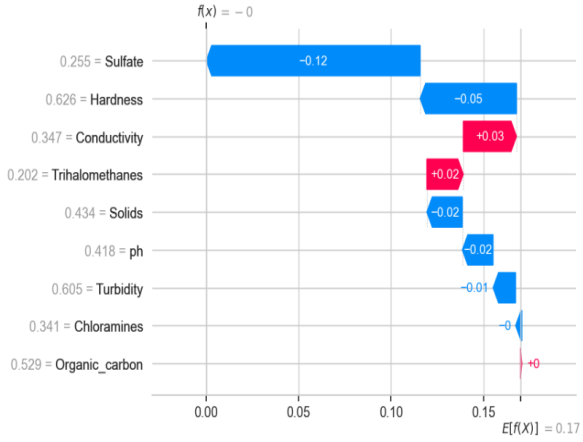

Fig. 10. SHAP Waterfall Graph

binary classifications as well. Counts from predicted and actual values is represented by confusion matrices. In confusion matrices True Negative is represented by “TN” it shows the number of negative examples which were labelled correctly. In the same way, True Positive is represented by “TP” it shows the number of positive samples which were labelled correctly. False Positive is represented by “FP” it shows the number of actual negative samples which were classified as positive. And False Negative is represented by “FN” it shows the number of actual positive samples classified as negative. For evaluation of WQI model we use Accuracy, Precision, Recall, Specificity, Mean Square Error, Sensitivity. These statistical measurements are as mentioned below:

- Accuracy=TP+TN/TP+FP+FN+TN
- Precision=TP/TP+FP
- Recall=TP/TP+FN
- F1-score=2*P*R/(P+R)

X. RESULT AND ANALYSIS

In this section, dataset used along with use of various Machine Learning algorithms for prediction is highlighted. The dataset is used is from Kaggle. This dataset consists of water quality metrics for 3276 different water bodies. Key features used to tally the results are: pH value, hardness, solids, Chloramines, Sulphates, Organic carbon, Trihalomethanes, Turbidity, Potability. The results of the different machine earning algorithms, namely, SVM, Decision Tree, Random Forest are discussed based on parameters like accuracy, precision, recall and F1-score. The main aspect taken into consideration is accuracy.

Accuracy=TP+TN/TP+FP+FN+TN

The results of the algorithms are mentioned below:

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my teacher, Nurjahan Nipa, for her invaluable guidance and unwavering

Fig. 11. Hardware Setup

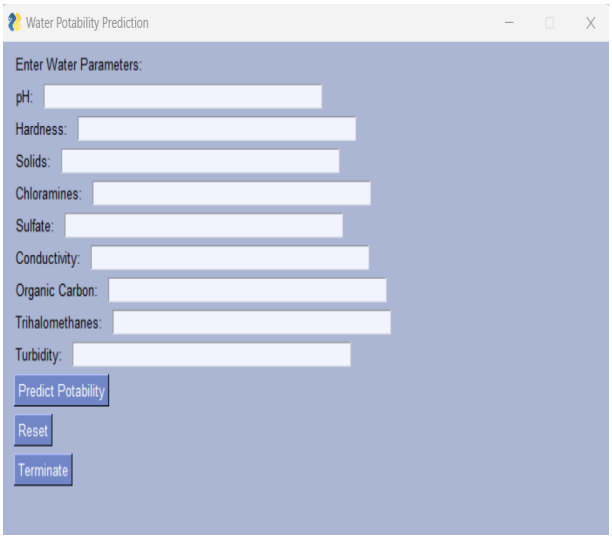

Fig. 12. User-dashboard

Fig. 13. Predicted-Output

TABLE II
MODEL ACCURACY

Moldel	Accuracy
KNN	66%
Logistic regression	63%
Naive Bayes	60%
Decision Tree	56%
SVM	67%
Decision Tree	65%

support throughout the course of this data science project. Her expertise, patience, and dedication were instrumental in helping me navigate the complexities of data analysis and machine learning.

Nurjahan Nipa's insightful feedback and constructive criticism played a pivotal role in shaping the success of this project. Her commitment to fostering a deep understanding of data science concepts and her willingness to go the extra mile to assist her students were truly commendable.

I would also like to extend my appreciation to my fellow classmates who collaborated with me on this project, as well as my friends and family for their constant encouragement.

XI. CONCLUSION AND FUTURE WORK

For the protection of the environment and the human health, Water quality prediction plays a very important role. With the advancement in the technology, Artificial and machine learning models can be used for the prediction of the same to make human life healthier and easier. In this paper, investigation of different Machine Learning algorithms on Water Quality Prediction dataset has been done. The simulation result shows that Random Forest algorithm outperforms the other two algorithms namely Support Vector Machine and Decision Tree. The validation on test dataset provides 68% accuracy which is 8% better than other algorithms. This clearly shows the effectiveness of the technique in the prediction of Water Quality. Accuracy can further be enhanced by training the model with larger number of samples.

REFERENCES

[1] https://www.geeksforgeeks.org/support-vector-machine-algorithm/, 2023.
[2] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning, 2023. Last accessed 16 September 2017.
[3] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm, 2023.
[4] https://www.javatpoint.com/machine-learning-naive-bayes-classifier, 2023. Last accessed 16 September 2017.
[5] https://www.javatpoint.com/machine-learning-random-forest-algorithm, 2023.
[6] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan, and Jose Garcia-Nieto. Efficient water quality prediction using supervised machine learning.
[7] Mohammed Al-Yaari, Hasan Alkahtani, and Mashael Maashi. Water quality prediction using artificial intelligence algorithms. 2020.
[8] Hritwik Ghosh, Mahatir Ahmed Tusher, Irfan Sadiq Rahat, Syed Khasim, and Sachi Nandan Mohanty. Water quality assessment through predictive machine learning. In Valentina Emilia Balas, Vijay Bhaskar Semwal, and Anand Khandare, editors, *Intelligent Computing and Networking*, pages 77–88, Singapore, 2023. Springer Nature Singapore.
[9] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie. Water quality prediction using machine learning methods. 2018.
[10] Kaggle. https://www.kaggle.com/datasets/adityakadiwal/water-potability, 2023.
[11] Manisha Koranga, Pushpa Pant, Tarun Kumar, Durgesh Pant, Ashutosh Kumar Bhatt, and R.P. Pant. Efficient water quality prediction models based on machine learning algorithms for nainital lake, uttarakhand. 2022.
[12] Navideh Noori, Latif Kalin, and Sabahattin Isik. Water quality prediction using swat-ann coupled approach. 2020.
[13] Jin Won Yu, Ju-Song Kim, Xia Li, Yun chol Jong, Kwang-Hun Kim, and Gwang-Il Ryang. Water quality forecasting based on data decomposition, fuzzy clustering, and deep learning neural network. 2022.