WATER QUALITY PREDICTION USING MACHINE LEARNING

Priya Kamble[1], Mithali Patil[2], Pallavi Patil[3], Sakshi Hanje[4] ,
Sejal Kamble[5], Vedika Hiraskar[6]

[1] *Priya Kamble, Data Science Department, DYPCET, Kolhapur, India*
[2] *Mithali Patil, Data Science Department, DYPCET, Kolhapur, India*
[3] *Pallavi Patil, Data Science Department, DYPCET, Kolhapur, India*
[4] *Sakshi Hanje, Data Science Department, DYPCET, Kolhapur, India*
[5] *Sejal Kamble, Data Science Department, DYPCET, Kolhapur, India*
[6] *Vedika Hiraskar, Data Science Department, DYPCET, Kolhapur, India*

ABSTRACT

This water quality analysis and prediction project focuses on utilizing machine learning algorithms, namely ***Random Forest, Decision Tree, and KNN, etc****. to analyze and predict water quality. The project involves collecting water quality data, preprocessing the data, applying exploratory data analysis techniques, implementing the algorithms, and evaluating their performance. Through this project, the objective is to analyze the collected water quality data, identify patterns and relationships between water parameters, and develop predictive models that accurately predict water quality based on those parameters. The models' performance will be evaluated using appropriate metrics, comparing their accuracy rates and identifying the most effective algorithm for water quality prediction. The project aims to contribute to sustainable water resource management by providing valuable insights into the factors influencing water quality. By accurately predicting water quality, it becomes possible to implement proactive measures, detect potential risks, and take timely actions to safeguard water resources. It promotes effective decision-making, and ensuring the well-being of ecosystems and communities that rely on clean water sources.*

Keyword - *Random forest, decision tree, KNN, and prediction, etc.*

1.INTRODUCTION

Water quality analysis and prediction play a crucial role in ensuring the safety and sustainability of our water resources. With increasing concerns about water pollution and its impact on human health and the environment, there is a growing need for accurate and reliable methods to assess and forecast water quality. In this context, machine learning algorithms such as Random Forest, Decision Tree, and KNN, etc. have emerged as powerful tools for water quality analysis and prediction. By utilizing a dataset that includes information on factors such as pH, temperature, dissolved oxygen, conductivity, and other relevant parameters, the project seeks to develop models that can accurately predict water quality. Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy. Decision Tree, on the other hand, constructs a treelike model of decisions and their consequences. KNN algorithm classifies samples based on the majority class of their k-nearest neighbors. The project involves several stages, including data collection, preprocessing, exploratory data analysis, algorithm implementation, and model evaluation. The collected data is carefully processed and cleaned to remove any outliers or missing values. Exploratory data analysis techniques are applied to gain insights into the relationships and patterns within the dataset. The Machine learning algorithms are then implemented to train models using the prepared data. The models' performance is evaluated using appropriate metrics, allowing for a comparison of their accuracy and suitability for water quality prediction. By accurately predicting water quality, this project aims to contribute to effective water resource management and environmental protection. The insights gained from the analysis can help

identify potential risks and guide decision-making processes related to water quality monitoring and remediation efforts.

2. LITERATURE SURVEY

In paper [1] authors proposed that this literature review underscores the importance of employing machine learning techniques for measuring water quality. By utilizing water quality parameters as feature vectors and employing classification algorithms, such as Decision Tree and K-Nearest Neighbor, the study demonstrates the efficacy of machine learning in predicting water potability. The findings contribute to the advancement of water quality assessment and provide valuable insights for future research and implementation in the field[1].

In paper [2] authors proposed a water quality prediction was generated for predicting if the water is safe to drink or not. This experiment was also conducted to compare the machine learning model performance between Decision Tree, Random Forest, XGBoost, and Logistic Regression to determine the most suitable technique for predicting Water Quality. The result of this experiment is XGBoost Algorithm gave the best accuracy at '71.23'% sitication algorithm, machine learning, water quality research[2].

3. SYSTEM DESIGN AND METHODLOGY

3.1 System Design

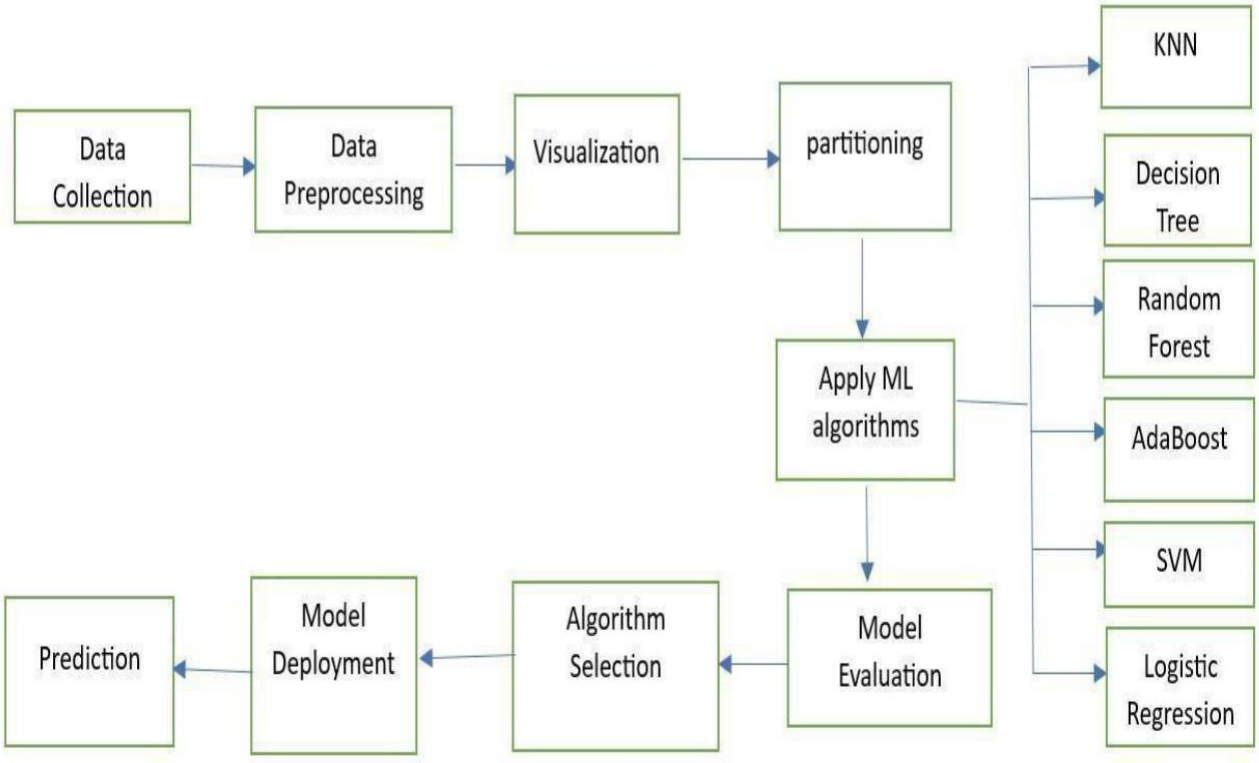

Fig -1: System Design

3.2 METHODOLOGY

Dataset:

This dataset includes pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability. These elements are present in water.

Importing libraries:

The required libraries such as pandas, matplotlib, and seaborn are imported. Loading the dataset: The dataset "water_potability.csv" is loaded using pandas' read_csv function and stored in the DataFrame variable df. The head() function is used to display the first few rows of the dataset.

Data Visualization:

Various visualizations are created using matplotlib and seaborn libraries to gain insights into the data distribution and relationships between variables. Visualizations include bar plots, histograms, pair plots, scatter plots, and correlation heatmaps.

Data Partitioning:

The dataset is divided into training and testing sets using the `train_test_split` function from the sklearn.model_selection module. Three different partitioning configurations are applied, with 20%, 30%, and 40% of the data reserved for testing, while the remaining data is used for training.

Model Building:

Two machine learning algorithms, K-Nearest Neighbors (KNN) and Decision Tree, are implemented using scikitlearn's , KNeighbors Classifier and DecisionTree Classifier classes, respectively. Additionally, a Random Forest model is also utilized. The models are trained on the training data and evaluated on the testing data. For each model, accuracy, precision, recall, and F1 score metrics are computed.

Accuracy Comparison:

The accuracy scores obtained from the three partitioning configurations are compared using bar plots.

Prediction on New Data:

The user is prompted to enter the values for various water attributes. The trained Decision Tree and Random Forest models are used to predict the potability of the water based on the provided attribute values.

Further Comparison and Evaluation:

Bar plots are created to compare accuracy, precision, recall, and F1-score among the KNN, Decision Tree, and Random Forest models. Modules: The modules in the water quality analysis and prediction project using Random Forest,

4. RESULTS

This research investigated how well machine learning approaches predicted the water quality elements of a water quality dataset. For this,, the most well-known dataset variables including Ph, Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity were acquired. Random Forest model perform well with an accuracy of 69%.

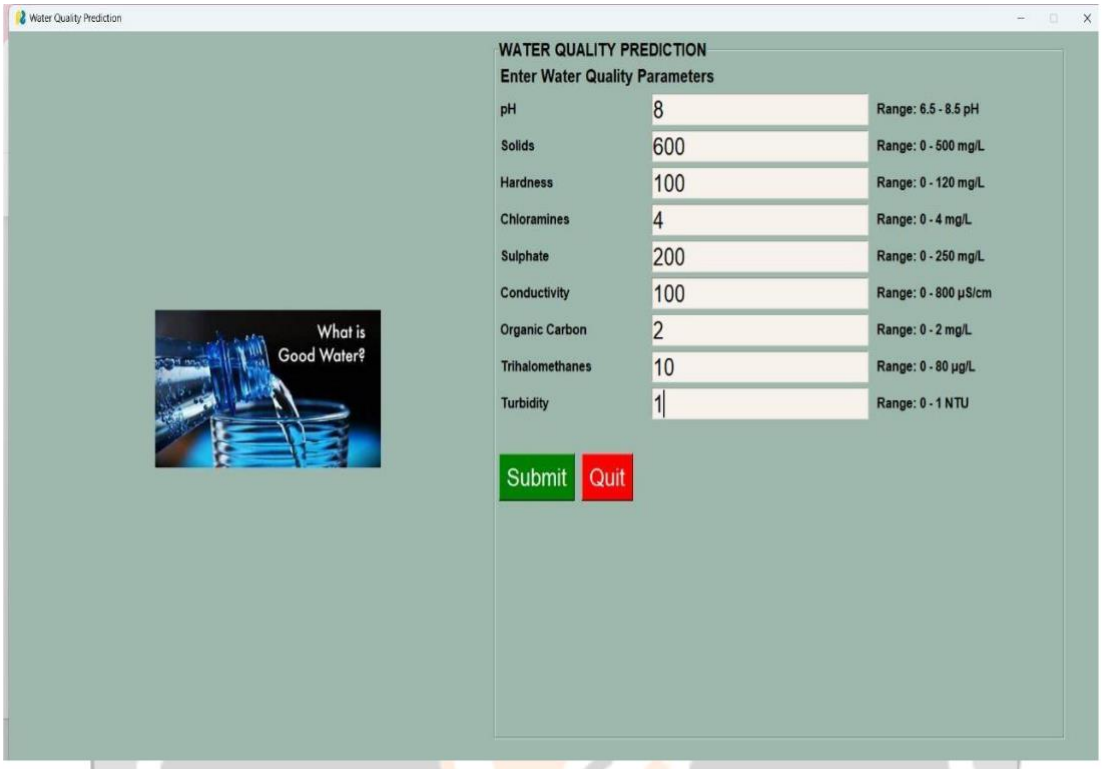

Fig -2: System Design

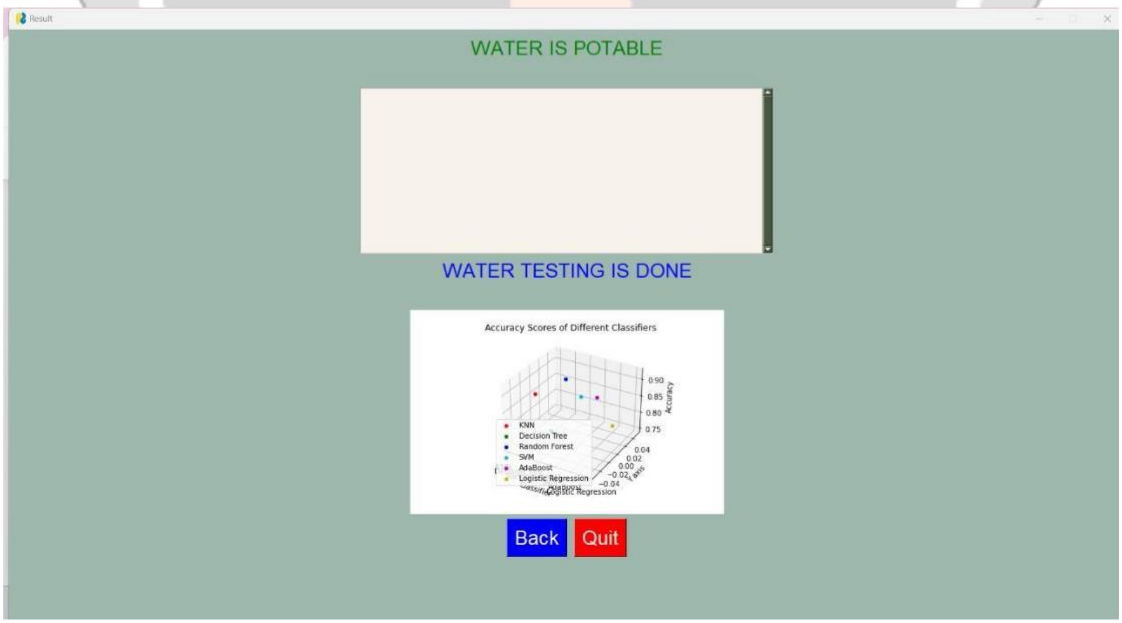

Fig -3: Water Potable

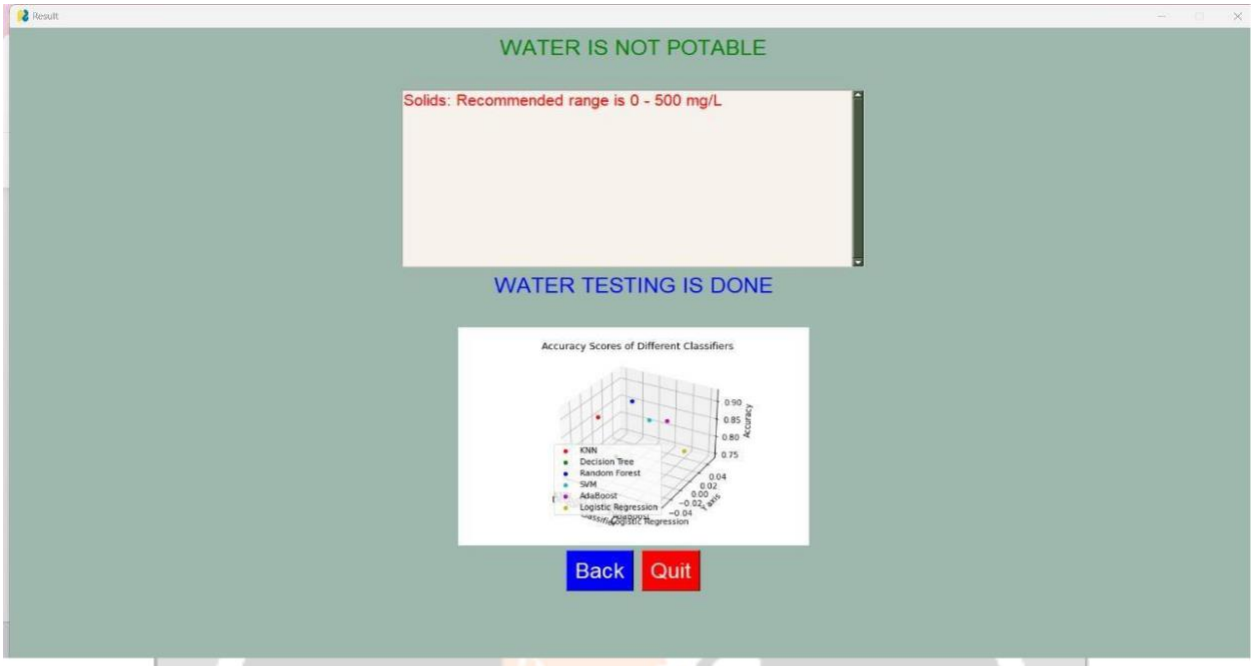

Fig -4: Water Not Potable

5. CONCLUSION

We are all aware of how vital water is to human health. Knowing the water's quality is crucial because if we consume water without first making sure it is safe to do so, we run the risk of getting sick. Numerous illnesses that are transmitted through water exist and if we consume non drinkable water, we risk contracting hazardous diseases. Consequently, the most crucial factor is understanding the water's quality. But this is where the real issue is. We must test the water at a lab, which is expensive and time consuming in addition to being necessary for determining the water's quality. In this study, we therefore provide a different strategy for predicting water quality using Machine Learning.

6. REFERENCES

[1]. Data-driven Water Quality Analysis and Prediction: A Survey(Gaganjot Kaur Kang ,Jerry Zeyu Gao ,Gang Xie)Published: 20 August 2020.

[2]. Water Quality Prediction Using Machine Learning(Sai Sreeja Kurra*1, Sambangi Geethika Naidu*2, Sravani Chowdala*3, Sree Chithra Yellanki*4, Dr. B. Esther Sunanda*5) Published:05May-2022

[3]. Li, Y., Linero, A.R., Murray, J.: Adaptive conditional distribution estimation with bayesian decision tree ensembles. Journal of the American Statistical Association, 1–14 (2022)

[4]. Rajaee, T., Boroumand, A.: Forecasting of chlorophyll-a concentrations in south san francisco bay using five different models. Applied Ocean Research 53, 208–217 (2015)