IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304043 |

Forecasting Water Quality: A Comparative Analysis of Predictive Models for Water Quality Assessment and Prediction

Pattan Shabbir Ali[1], Palaparthi Gopi[2], Parre Hosanna[3], Myla Srinu[4], M.Hanumantha Rao[5]

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India[1,2,3,4]

Assistant Professor, Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India[5]

ABSTRACT: Maintaining water quality is paramount for environmental and human health. However, its dynamic nature necessitates robust forecasting methods. This review paper focuses on the application of machine learning models for water quality assessment and prediction. We conduct a comparative analysis, emphasizing the strengths and weaknesses of various algorithms like Decision Tree Classifiers and K-Nearest Neighbors (KNN) in capturing complex water quality data.

The analysis explores the integration of Exploratory Data Analysis (EDA) techniques for data preprocessing and partitioning strategies for model training and validation. We delve into the optimization of KNN models through hyperparameter tuning, aiming to enhance their accuracy and generalizability. This comparative approach highlights the trade- offs between interpretability (e.g., Decision Trees) and complex pattern recognition (e.g., KNN) when selecting machine learning models for water quality forecasting.

KEYWORDS: Water Quality, Explorarory Data Analysis, KNN, Partitioning, Decision Tree Classifier.

I. INTRODUCTION

Maintaining good water quality is a critical concern for ensuring the health of ecosystems and human populations. However, water quality is constantly in flux due to a multitude of factors, including natural variations, pollution events, and climate change. This necessitates the development of reliable forecasting methods to anticipate these changes and enable proactive water management strategies. This review paper explores the application of machine learning models for water quality assessment and prediction. We present a comparative analysis of various machine learning algorithms, focusing on their effectiveness in capturing the complexities inherent in water quality data.

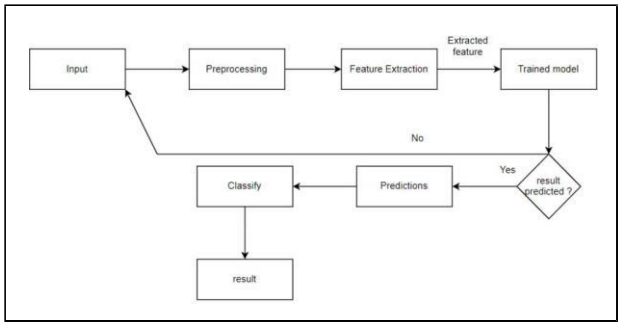


Fig1. System Architecture

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304043 |

II. LITERATURE REVIEW

1.) Performance Analysis of Machine Learning Techniques for Predicting Water Quality Index using Physiochemical Parameters Sazia Tabassum, CB Kotnala, Raj Kumar Masih, Mohammed Shuaib, Shadab Alam,TM Alar etc..,In this project , a machine learning-based model for predicting WQI based on physicochemical parameters was proposed, which overcomes the challenge of capturing complex, non- linear relationships between physicochemical parameter and water quality.

2.) Water Quality Prediction Using Machine Learning Annaji Kuthe, Chaitanya Bhake, Vaibhav Bhoyar, Aman Yenurkar, Ketan Gawale etc..,In this project , a water quality classification model using Back Vector Machine (SVM) and Extraordinary Gradient Boosting (XGBoost) was proposed to predict the water quality.

3.) Efficient Drinking Water Quality Analysis using Machine Learning Model with Hyper-Parameter Tuning
M. Maheswari, M Arthy, V. Samuthira Pandi etc..,The research paper by Osim Kumar Pal (2022) focuses on efficient drinking water quality analysis using machine learning models like Decision Tree and Random Forest with hyper-parameter tuning.In this project, the authors examined a variety of supervised machine learning algorithms in order to identify the quality of the water, including Random Forest (RF) and Decision Tree (DT) algorithms.

4.) Efficient Data-Driven Machine Learning Models for Water Quality Prediction
Elias Dritsas, Maria Trigka etc..,In this project , a supervised learning approach was used to design as accurate as possible predictive models from a labelled training dataset for the identification of water suitability, either for consumption or other uses.

5.) Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability
Mustafa Yurtsever, Murat Emeç etc..,In this project, a dataset obtained from the Kaggle website was used to classify water quality in real-time efficiently and quickly, and the results showed that the SVR XGboost Hybrid (SXH) model had the best performance ratio (Accuracy, F1-score).

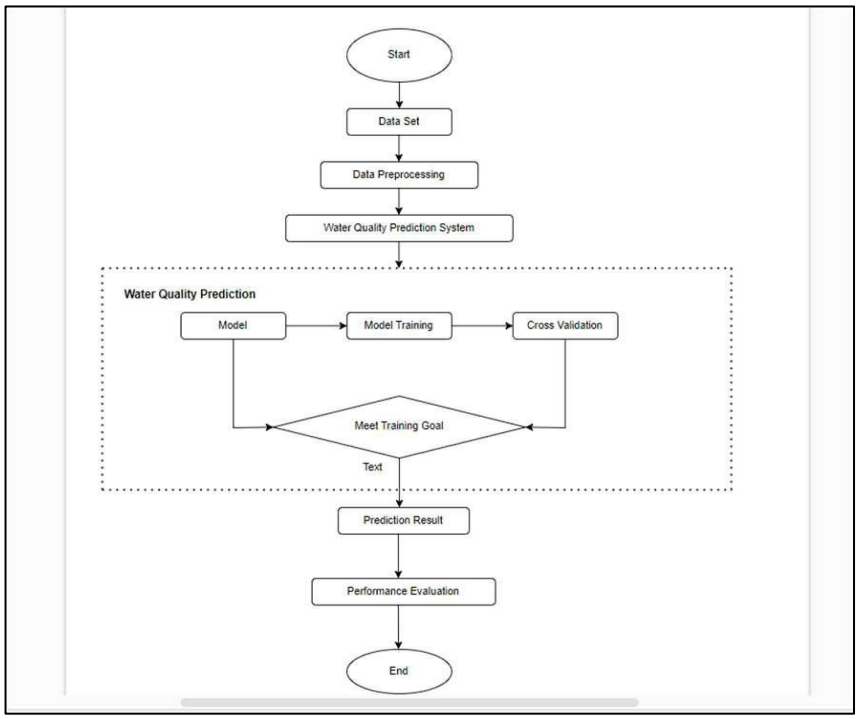
III. METHODOLOGY

The data analysis process initiates by acquiring a dataset that is pertinent to water quality measurements. This dataset is crucial as it forms the foundation for all subsequent analyses and model training. Once the data is obtained, the next step involves data preprocessing. During this phase, the dataset undergoes various transformations and refinements to ensure its suitability for analysis.Common preprocessing tasks include error correction, filling in missing values, and formatting the data into a standardized structure. This step is essential as it ensures that the data is clean, consistent, and ready for further exploration and modeling

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

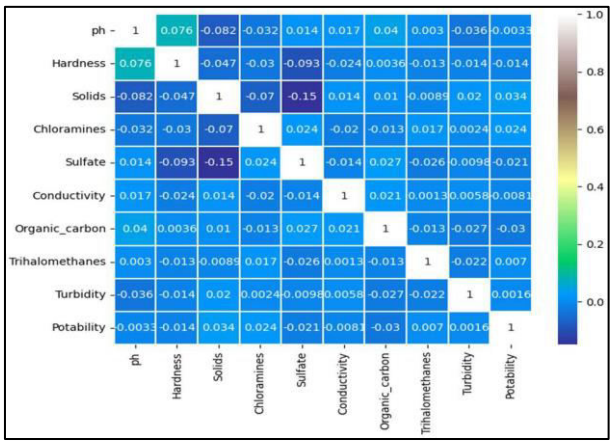
| DOI:10.15680/IJIRSET.2024.1304043 |



Following data preprocessing , the next phase is data analysis.In this stage, various analytical techniques are applied to the preprocessed data to uncover patterns,trends and relationships.Exploratory analysis methods are often used to gain insights into the underlying structure of the data and identify potential variables of interest.

After completing data analysis, the next steps involve model training and cross-validation.The dataset is divided into two subsets: a training set used to train the model and a validation set used to assess the model's performance.This iterative process may involve trying out multiple models and selecting the one that performs best on the validation set, thus optimizing the predictive power of the model such that a suitable model is selected based on performance evaluation.

Data Cleaning and Exploration: Effective data cleaning techniques are used to handle missing values,outliers and data inconsistencies present in the water quality dataset.



Using EDA techniques we can identify potential patterns, relationships,and anomalies.Data visualization techniques, such as scatter plots,box plots, and heatmaps will be employed to visually explore the relationship between water

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304043 |

quality parameters.It aims to ensure the quality and integrity of the input data, therby enhancing the performance and reliability of the predictive models for water quality assessment and forecasting.

Decision Tree Classifier : The project focuses on implementing and evaluating a Decision Tree Classifier for water quality prediction using the scikit-learn library. It begins by importing essential modules, such as DecisionTreeClassifier from sklearn.tree, and evaluation metrics like accuracy_score, confusion_matrix, and precision_score from sklearn.metrics. The Decision Tree Classifier is initialized with specific parameters, including the gini criterion for split quality measurement, a minimum sample threshold for internal node splitting (min_samples_split=10), and the best strategy for selecting node splits (splitter='best'). This setup lays the foundation for subsequent evaluation and optimization of the classifier in the water quality prediction task.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score,confusion_matrix,precision_score
dt=DecisionTreeClassifier(criterion= 'gini', min_samples_split= 10, splitter= 'best')
dt.fit(X_train,Y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(min_samples_split=10)
```

K-Nearest Neighbors : The process begins with importing KNeighborsClassifier from scikit-learn to implement the K-Nearest Neighbors algorithm. An instance of the classifier is initialized with 'manhattan' distance metric and 22 neighbors. Training ensues with the fit() method on training data (X_train, Y_train), involving distance calculation and storing K nearest neighbors per data point.

```
from sklearn.neighbors import KNeighborsClassifier

knn=KNeighborsClassifier(metric='manhattan', n_neighbors=22)
knn.fit(X_train,Y_train)
```

```
KNeighborsClassifier
KNeighborsClassifier(metric='manhattan', n_neighbors=22)
```

Predictions on test data (X_test) are made using predict() method, and results are stored in prediction_knn. Model accuracy on test set is computed with accuracy_score(Y_test, prediction_knn) and printed. Furthermore, confusion_matrix(prediction, Y_test) is utilized for a detailed overview of correct and incorrect predictions. While KNN is effective for handling non-linear relationships, its performance depends on factors like distance metric, K value, and data quality.

IV. EXISTING SYSTEM

The existing system for water quality forecasting relies on conventional methods like manual sampling and lab analysis, which are time-consuming and may not capture real-time variations. These methods lack spatial and temporal coverage and may not promptly address emerging water quality challenges. To overcome these limitations, the comparative analysis of predictive models explores modern machine learning algorithms efficacy in forecasting water quality, offering real-time insights and adaptability to changing conditions. By analyzing large datasets, these models can identify complex patterns and relationships, enabling more accurate predictions. This shift towards predictive modeling allows for proactive water quality management, facilitating early detection of contamination events and preventive measures to safeguard public health and the environment.

V. PROPOSED SYSTEM

It aims to develop and evaluate various predictive models for forecasting water quality parameters. The system utilizes historical data on water quality parameters, such as pH, dissolved oxygen, biochemical oxygen demand, and other relevant factors, to train and evaluate different machine learning models like “Quadratic Discriminant Analysis,K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier”.

The proposed system works on a range of predictive models, including decision trees classifier, KNN and Model Optimization.These models are trained on the historical water quality data, with appropriate preprocessing and

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304043 |

feature selection techniques applied. By leveraging historical data and advanced machine learning techniques, the system can provide valuable insights and predictions for water quality management.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
qda	Quadratic Discriminant Analysis	0.6741	0.6820	0.3390	0.6556	0.4459	0.2475	0.2747	0.0550
rf	Random Forest Classifier	0.6658	0.6707	0.3162	0.6405	0.4221	0.2243	0.2521	1.3520
et	Extra Trees Classifier	0.6649	0.6686	0.2958	0.6484	0.4046	0.2160	0.2481	0.3470
gbc	Gradient Boosting Classifier	0.6477	0.6312	0.2437	0.6234	0.3483	0.1663	0.2016	0.8810
lightgbm	Light Gradient Boosting Machine	0.6424	0.6745	0.4035	0.5550	0.4662	0.2078	0.2142	0.9630
xgboost	Extreme Gradient Boosting	0.6349	0.6538	0.4261	0.5376	0.4747	0.2010	0.2044	0.1940
nb	Naive Bayes	0.6187	0.5827	0.2086	0.5256	0.2974	0.0984	0.1189	0.0540
ridge	Ridge Classifier	0.6121	0.0000	0.0011	0.1000	0.0022	0.0014	0.0083	0.0640
lda	Linear Discriminant Analysis	0.6121	0.4960	0.0011	0.1000	0.0022	0.0014	0.0083	0.0330
lr	Logistic Regression	0.6116	0.4977	0.0011	0.0500	0.0022	0.0005	0.0021	1.0290
dummy	Dummy Classifier	0.6116	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0280
ada	Ada Boost Classifier	0.6028	0.5449	0.2086	0.4750	0.2883	0.0686	0.0800	0.6140
dt	Decision Tree Classifier	0.5878	0.5678	0.4783	0.4699	0.4732	0.1350	0.1354	0.0980
knn	K Neighbors Classifier	0.5605	0.5196	0.3004	0.4092	0.3457	0.0272	0.0278	0.1770
svm	SVM - Linear Kernel	0.4972	0.0000	0.4943	0.2260	0.2812	-0.0053	-0.0221	0.0850

VI. EXPERIMENTS AND RESULTS

The user interfaces in this water quality prediction system are designed to utilize machine learning models, such as Decision Tree Classifiers or KNN algorithms, to assess water potability based on input parameters like pH, hardness, and chloramines. Decision Tree Classifiers segment the input space recursively, while KNN relies on the majority class of nearby data points for classification. These models undergo rigorous preprocessing and hyperparameter tuning to ensure accuracy, including handling missing values and optimizing parameters like maximum depth and number of neighbors.

After preprocessing and optimization, the trained models are deployed via user interfaces for real-time predictions. Users input water quality parameters, and the models classify the water as "Healthy" or "Unhealthy" based on learned decision rules or distance calculations. These predictions guide decisions regarding water treatment and distribution, contributing to the goal of ensuring safe and clean water resources.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7018	0.6907	0.3708	0.7333	0.4925	0.3122	0.3486
1	0.6476	0.6924	0.2386	0.6176	0.3443	0.1635	0.1981
2	0.6960	0.6900	0.3977	0.6863	0.5036	0.3062	0.3299
3	0.6872	0.6325	0.3523	0.6889	0.4662	0.2763	0.3074
4	0.6652	0.6703	0.3295	0.6304	0.4328	0.2271	0.2512
5	0.7048	0.6840	0.3750	0.7333	0.4962	0.3171	0.3528
6	0.6432	0.6549	0.3182	0.5714	0.4088	0.1819	0.1979
7	0.6608	0.7058	0.3068	0.6279	0.4122	0.2116	0.2384
8	0.6828	0.7046	0.3636	0.6667	0.4706	0.2711	0.2965
9	0.6520	0.6944	0.3371	0.6000	0.4317	0.2084	0.2264
Mean	0.6741	0.6820	0.3390	0.6556	0.4459	0.2475	0.2747
Std	0.0220	0.0219	0.0426	0.0522	0.0472	0.0534	0.0567

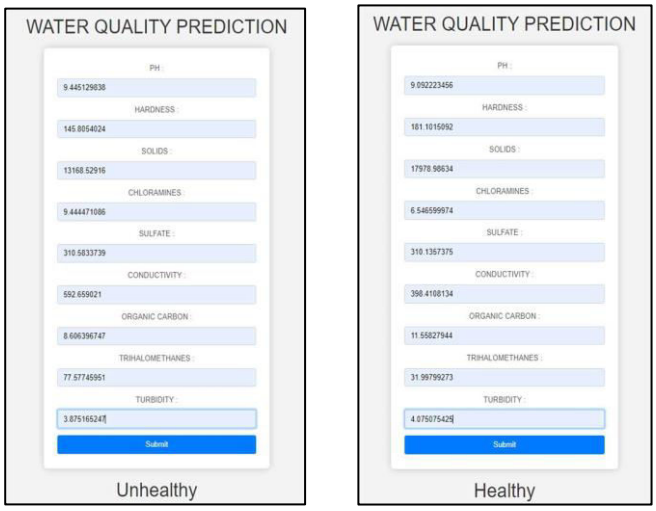
QuadraticDiscriminantAnalysis

```
QuadraticDiscriminantAnalysis(priors=None, reg_param=0.0,
                              store_covariance=False, tol=0.0001)
```

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304043 |



VII. CONCLUSION

This project conducted a thorough examination of water quality prediction utilizing diverse machine learning techniques. Initial exploratory data analysis unveiled dataset characteristics, including shape, missing values, and class distributions. Visualization methods such as histograms and correlation heatmaps offered insights into feature distributions and relationships. Decision Tree (DT) and K-Nearest Neighbors (KNN) classification models were constructed and refined through hyperparameter tuning to enhance predictive accuracy. The project also addressed outlier detection and removal to bolster model robustness. Leveraging the PyCaret library facilitated streamlined model comparison, tuning, and deployment, aiding efficient model selection and implementation. Overall, the project aimed to develop precise and dependable predictive models for water quality assessment and forecasting, thereby supporting environmental monitoring and management initiatives.

REFERENCES

1.)Sazia Tabassum, CB Kotnala, Raj Kumar Masih, Mohammed Shuaib, Shadab Alam, Tariq Mousa Alar (2023) “Performance Analysis of Machine Learning Techniques for Predicting Water Quality Index Using Physiochemical Parameters” - International Conference on Sustainable Computing and Smart Systems (ICSCSS) .
2.) Annaji Kuthe, Chaitanya Bhake, Vaibhav Bhoyar, Aman Yenurkar, Ketan Gawale (2023) “ Water Quality Prediction Using Machine Learning “ - International Journal of Computer Science and Mobile Computing, Vol.12 Issue.4, pg. 52-59.
3.)M. Maheswari, M Arthy, V. Samuthira Pandi (2023) “Efficient Drinking Water Quality Analysis using Machine Learning Model with Hyper-Parameter Tuning “ - 7th International Conference on Intelligent Computing and Control Systems (ICICCS).
4.)Elias Dritsas, Maria Trigka (2023) “Efficient Data-Driven Machine Learning Models for Water Quality Prediction ”.
5.)Mustafa Yurtsever, Murat Emeç “Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability” - Ege Akademik Bakış Dergisi -Cilt: 23 Say: 2.
6.)Annaji Kuthe, Chaitanya Bhake, Vaibhav Bhoyar, Aman Yenurkar, Vedant Khandekar, Ketan Gawale (2022) “Water Quality Analysis Using Machine Learning” - International Journal for Research in Applied Science and Engineering Technology(IJRASET).
7.)Raavi Akshay, Gadiraju Tarun, Pinapothini Uday Kiran, K. Durga Devi, M S Vidhyalakshmi “Water-Quality-Analysis using Machine Learning”.
8.)N. Nasir, Afreen Kansal, Omar Alshaltone, F. Barneih, Mustafa Sameer, Abdallah Shanableh, Ahmed Al-Shamma'a (2022) “Water quality classification using machine learning algorithms” - Journal of Water Process Engineering
9.)Mohamed Torky, Ali Adam Bakhiet, Mohamed Bakrey, Ahmed Adel Ismail, Ahmed I. El Seddawy (2024) “Recognizing Safe Drinking Water and Predicting Water Quality Index using Machine Learning Framework” - International Journal of Advanced Computer Science and Applications(IJACSA), Volume 14 Issue 1, 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details