TEAM – 22

Water Quality Prediction Using Machine Learning Algorithm

Pradeep Kumar Gupta, Amrita Vishwa Vidyapeetham, Bengaluru, India
Kartthikeyan N, Amrita Vishwa Vidyapeetham, Bengaluru, India.
Mukil LD, Amrita Vishwa Vidyapeetham, Bengaluru, India.

Abstract: This study attempts to address an essential issue of providing safe and good quality water that is necessary for public health and environment preservation. We use machine learning algorithms for carrying out the project "water quality prediction" which has a dataset containing diverse water quality parameters such as pH, Iron, nitrate, chloride, zinc, etc., which we intend to classify and predict the water quality. Interestingly, the use of SVM, Decision Tree, KNN, XGBoost, and Random Forest Classifier algorithms; and most importantly random forest classifier gives about 87 percent prediction accuracy for the target attributes. Our methodology ensures predictive accuracy by meticulously processing the dataset preprocesses, such as dealing with missing values as well as standardization of features. This study does not only demonstrate the Random Forest Classifier supremacy but also evaluates different machine learning algorithm performances to provide knowledge on water quality assessment. This study serves as a basis for detecting water-quality problems early on, which can aid in preventive measures and resource management when dealing with water resources.

Keywords: water quality prediction, random forest classifier, comparative analysis, predictive accuracy.

1. INTRODUCTION

Safe, clean water is of paramount importance in ensuring population health and sustaining an ecosystem. It is in light of this need that our research aims at tackling the multiple dimensions of accessing clean water in terms of public health as well as eco-friendliness. Our project, "Water Quality Prediction," utilizes sophisticated machine learning methods to increase knowledge about water quality parameters.

Our study is founded on a comprehensive core dataset involving several essential water quality indices such as pH, iron, nitrate, chloride, and zinc, among others. Complexity notwithstanding, our prime objective is determining and forecasting water quality with respect to these parameters. To maintain safe and wholesome water sources, there must be a sense of urgent nature in carrying out this task.

The methodology employed in this project features several machine learning algorithms, namely SVM, Decision Tree, KNN, XGBoost, and the noteworthy Random Forest Classifier. During our empirical exploration, the best-performing model was the Random Forest Classifier which had a prediction accuracy rate of 87% for the target attributes. The robust algorithmic features make this a credible predictive model for water quality in decision-making or allocation of water resources management operations.

However, we ensure that our approach succeeds by processing data very carefully where we take care of missing values and features' normalization to achieve reliable predictive models. Our study is not just another addition to the already expanding field of water quality forecasting, as it does more than just demonstrate the dominance of the Random Forest Classifier.

In addition, our study forms a fundamental basis for detecting water problems in the early stages. Identification of children at an early stage promotes informed decision-making. Our project in short seeks to connect the dots between the data-driven insights and the actionable strategy that if followed will be more adaptive and responsible in managing water resources to become more resilient in the long run. As for the next part, the details of our methods, results, and discussion are provided in this section, giving an in-depth look into the connection between machine learning and predicting water quality. Keywords: Random Forest Classifier as a water quality prediction tool using machine learning algorithms: comparative analysis and predictive accuracy

2. LITERATURE REVIEW

The study looks at [1] a big problem of getting worse water quality in places like ponds, lakes and rivers. This is happening because factories are putting wastewater into these waters without cleaning it first and mixing it with sewage. Old ways to check water quality take a long time, so people are looking into using Artificial Intelligence (AI) with AutoDL as a fix. AutoDL, a part of AutoML, makes the deep learning process automatic. This cuts down on human work and helps more people use it easily. The study wants to see how AutoDL works with regular deep learning models in checking water quality. It will focus on things like Total Dissolved Salts, pH, Phosphate and Turbidity. The reason for this study is that social well-being and technology progress are very important. It takes care of how clean water affects people's health and needs a big comparison study using artificial intelligence models. The writers want to check out what automated handling can do, while also knowing its shortcomings and connection to society. This study is important for showing the difference between old deep learning methods and automatic models. It helps us understand how well they work and their settings when looking at water quality.

This study [2] looks at how water quality is spread out in the Taihu Lake area. It uses a random forest regression (RFR) to figure out three important water quality things. By adding 16 watershed parts, RFR makes maps that show how water quality changes in different locations. This helps to see differences between places more clearly. The study looks at the problems with normal watch stations and shows how useful models like RFR are for guessing water quality using things about watersheds. The Shapley Additive method (SHAP) is used to understand what causes changes in water quality. The study's information is important for looking after water quality and making things better in Taihu Lake area. This gives a fact-based way to improve the situation.

The study [3] is about using machine learning methods to guess water quality levels. It focuses on total dissolved solids (TDS) and electrical conductivity (EC). These include multi-layer perceptron neural networks, support vector machines, and decision trees for predicting the results of a test or measurement. The data has 372 monthly readings and eight things to predict. Team learners, like adaptive boosting, bagging in ensemble learning, and random forest (RF) help improve single models. The findings show that group learners make the single models better. RF and DT models

are very accurate. The research shows that machine learning can help predict water quality. It also stresses that group models are better than single ones. The research uses a method called K-Fold cross-validation to check the testing data. It also uses different measurement tools for comparing group models with single ones.

The research suggests [4] a better Water Quality Index (WQI) model for safe assessment of coastal water quality, mainly focused on Cork Harbour. The study compares eight popular machine learning (ML) methods, like Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGB), Extra Tree (ExT), Support Vector Machine (SVM), Linear Regression (LR) and Gaussian Naïve Bayes Model for guessing The information is divided into learning and checking parts, checked fully by ten methods of group cross-checking. Performance measures such as RMSE, MSE, MAE, R2 and PERI are used to evaluate its performance. The research shows that tree-based methods like DT, ExT and group tree models like XGB and RF are much better than others. This suggests they work well when it comes to making guesses about water quality indexes using less uncertain ways.

The study recommends [5] a new method for testing water quality called WQI and grouping system to get the right results. It mostly focuses on Cork Harbour. The study uses four machine teaching ways, like Support Vector Machines (SVM), simple-minded Bayes (NB), Random Forest (RF) and K-Nearest Neighbor (KNN). It also uses a method called XGBoost to help with this. These assist in accurately categorizing water quality. The KNN and XGBoost techniques are better than others at predicting water quality categories. The XGBoost classifier is the best among all. The study also introduces new WQI models (WQM and RMS) and a standard method for grouping things to make testing water quality better. The study offers helpful advice on how to measure and group water quality using computer learning techniques and new ways of sorting.

This study is all about [6] the big need to watch over water quality for drinking and using machine learning methods to find pollution getting into systems that deliver water. The study checks how well two famous sorting methods, Decision Tree and Support Vector Machines (SVM), work with an actual set of data from a water treatment plant in Tunisia. The goal is to pick the best method for checking water quality. The study shows how important it is to always watch water quality, especially in smart cities. Also, machine learning can help solve this problem. The check is made using exactness, neatness, and return measures to compare how Decision Tree and SVM methods work in this situation.

This study focuses [7] on classifying water quality into four status categories: In good condition, a bit dirty, okay level of filth and very messy. These are important for finding out how to use and take care of water sources rightly. The study uses two ways of sorting, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), to check if their groupings for water quality are right. This is done by using different steps. The goal is to figure out which way works best for classifying water quality levels using 10-step Cross-Confirmation. The study shows that SVM, especially with a line-based kernel, is better than KNN. Its precision is 92.40%, but KNN's is just 71.28%. The

study shows how important it is to test water quality regularly. It also shows that SVM works the best in this situation.

SVM Result

No	Kernel	F Measure				Mean
		Good Condition	Lightly Polluted	Medium Polluted	Heavily Polluted	
1	Radial	12.50%	18.20%	23.50%	41.50%	**23.93%**
2	Linear	92.10%	84.20%	93.30%	100.00%	**92.40%**
3	Polynomial	84.80%	76.40%	78.90%	75.00%	**78.78%**
7	Sigmoid	27.30%	14.30%	27.30%	24.00%	**23.23%**

This study [8] looks at the important problem of worsening water quality in nature and the need for good ways to predict and handle this. The research uses Artificial Neural Networks (ANN) and time series analysis. It makes a predictive model with past water quality data from 2014, gotten from the United States Geological Survey (USGS). We look at four main water quality measures. These are chlorophyll, specific conductance, dissolved oxygen and cloudiness. The aim is to make models that guess water quality numbers by using their present values. The study tries to make water management methods better, dependable and easy-to-use. It may help improve future guessing or checking of water quality.

This study [9] is about sorting water quality, a very important part of managing water resources. To make this faster, we use special sorting systems like Support Vector Machine (SVM), Probabilistic Neural Network (PNN) and K-Nearest Neighbor (KNN). The study wants to see how these methods work in sorting water quality. It will measure their performance using error rate and error amount as judging tools. SVM is the best because it makes no mistakes, while KNN has the worst results. This is shown by how many and how big those errors are. This study shows that SVM and PNN are good ways to classify water quality.

This study [10] focuses on the need for good water quality forecast in Johor River Basin. It's getting worse because of different human actions. The research uses different methods for models, like Adaptive Neuro-Fuzzy Inference System (ANFIS), Radial Basis Function Neural Networks (RBF-ANN), and Multi-Layer Perceptron Neural Networks (MLP-ANN). To reduce noisy data, it's suggested to use a Wavelet De-noising Technique based on Neuro-Fuzzy Inference System (WDT-ANFIS). We use three ways to test how well the models are working. The field data uses measurements for water quality like amount of nitrogen, dirt in the water and ph level. The findings show that WDT-ANFIS is better than other models. In Scenario 2, it shows more accurate results. This study is very important to manage water resources well in that area.

This research studies [11] the urgent need to guess a feature called Water Quality Index (WQI) for 1944 wells in Vellore district. It uses something like setting up fake nerves, or just 'Artificial Neural Networks' with three different ways of working - Tanh, Maxout and Rectifier. The study uses 15 water under the ground data from 2008 till 2017 to guess WQI. If these factors fit certain levels, the water is seen as good for drinking. Techniques for preparing and extracting features are used on the data. The new part of this study is the mixing of three functions to predict WQI. We

compare the new method to others and see how well it works. Methods and algorithms employed: Artificial Neural Networks (ANN), Tanh, Maxout and Rectifier are types of activation functions used in computers to make them work better.

This paper talks [12] about the very important job of checking water quality, mainly for drinking. It also considers how our environment is being affected and looks at growing issues caused by waste from factories and pollution. It uses both Statistical Quality Control (SQC) methods and Bayesian ways to improve the accuracy of classifications. The way of doing it is by collecting data from sensors, looking at a chart to see if the values are within safe limits, and then using an easy-to-use math formula called Naive Bayesian algorithm for predicting water quality. The new method wants to cut down on mistakes that happen when manual testing for chemicals is done. The article study talks about different ways to check and rate water quality. However, the plan we suggest mainly looks at classifying water types based on past data and regular measurements.

3. PROPOSED METHODOLOGY

Many businesses are now using Machine Learning methods to understand and predict how materials behave. In this study, we use machine learning methods like Neural Networks (ANN), Decision Trees (DT), and Random Forests (RF). We also look at Multiple Linear Regression (MLR) and Support Vector Machines (SVM) to predict the Water Quality Index. Specifically, these techniques are used for forecasting Total Dissolved Solids in water (These ways are chosen because they're used a lot, and have shown that they can predict results well in similar studies and important data mining methods. After that, group learning is used to make models for TDS and EC. The main picture, called Fig in short, shows how the method works. It includes using group learning as a big part of it to make better predictions.

3.1 Data Preprocessing

The preprocessing methodology employed in this study involves a systematic series of steps aimed at refining the raw water quality dataset for subsequent analysis and modeling. Initially, a subsampling approach is implemented to curate the dataset, ensuring a manageable size for downstream computations. The decision to retain 352,250 instances strikes a balance between maintaining data richness and computational feasibility. The random sampling procedure is executed using the Pandas library, resulting in a more manageable dataset that preserves the inherent variability of the original data.

Subsequently, irrelevant attributes are systematically identified and removed from the dataset. Attributes such as 'Index,' 'Lead,' 'Source,' 'Water Temperature,' 'Air Temperature,' 'Month,' 'Day,' and 'Time of Day' are deemed extraneous to the analysis and are consequently dropped. This attribute removal process streamlines the dataset, focusing on key features relevant to water quality without introducing noise or unnecessary complexity. The efficacy of this step is validated by ensuring that each attribute to be removed exists in the dataset before execution.

To address the issue of scientific notation in specific columns, particularly 'Iron' and 'Manganese,' a formatting step is introduced. This conversion ensures that values originally represented in scientific notation are reformatted to a more conventional, human-readable form. Furthermore, categorical data, exemplified by the 'Color' attribute, is transformed into a numeric format using label encoding. This facilitates the inclusion of such attributes in machine learning models that require numerical input. Following these transformations, additional measures are taken to convert 'Iron' and 'Manganese' columns to numeric (float) types, fostering consistency in the dataset's data types.

Handling missing values is a critical aspect of data preprocessing. In this study, missing values in key columns, including 'pH,' 'Iron,' 'Nitrate,' and others, are imputed using mean values derived from their respective columns. This approach ensures a balanced representation of the dataset and avoids introducing bias by replacing missing values with the mean of the entire dataset. Finally, the preprocessed dataset is saved to a new CSV file, 'modified_WQP_dataset.csv,' capturing all the applied transformations. This refined dataset is now poised for further exploratory analysis and the development of predictive models, providing a robust foundation for subsequent stages of the study.

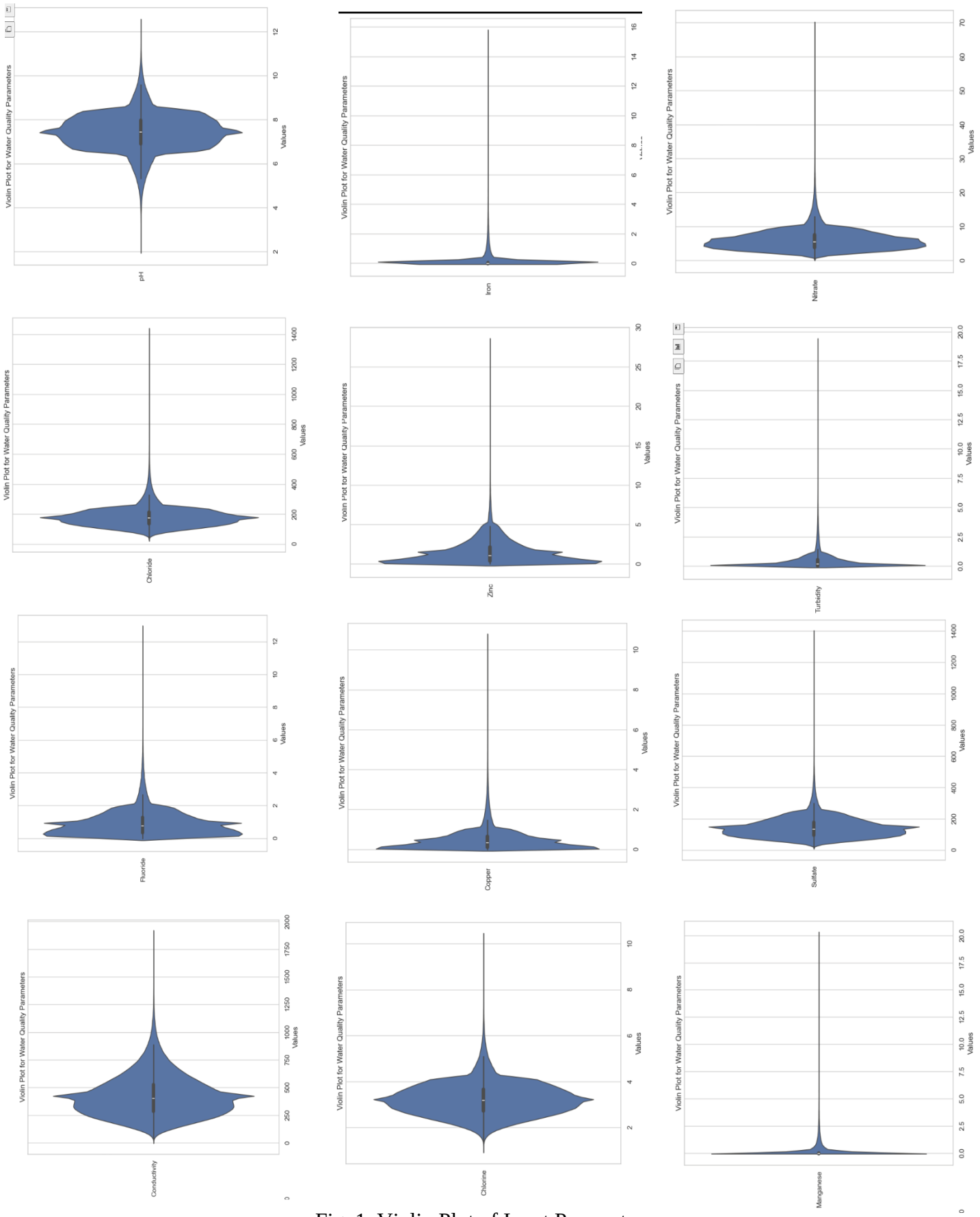

Fig. 1. Violin Plot of Input Parameters.

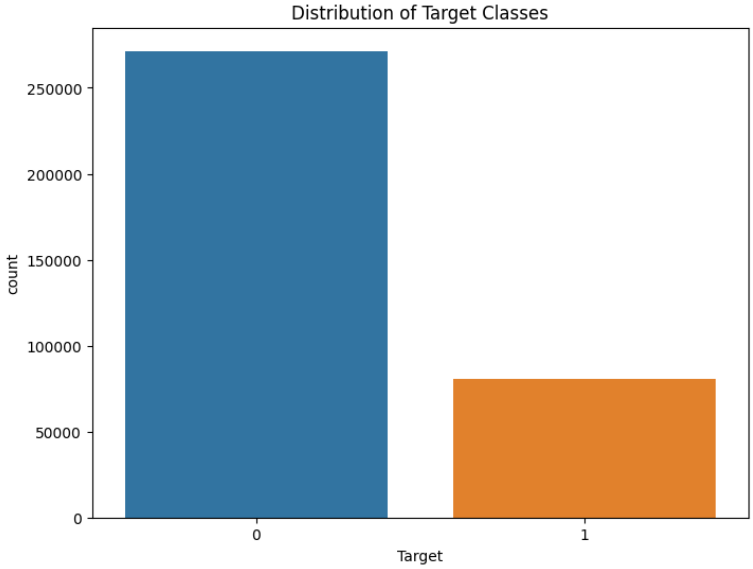

Fig. 2. Distribution of Target Attribute

3.2 Methodologies or Algorithms

3.2.1 Random Forest Classifier:

The Random Forest Classifier, a prominent ensemble learning algorithm, played a pivotal role in our water quality prediction project. Known for its versatility and robustness, this classifier demonstrated an impressive 87% prediction accuracy for the target attributes in our dataset. By aggregating predictions from multiple decision trees, the Random Forest Classifier minimizes overfitting and enhances predictive performance. Its ability to handle diverse data types, including categorical and numerical features, makes it particularly suitable for our comprehensive water quality dataset. Moreover, its interpretability and feature importance analysis provide valuable insights into the crucial factors influencing water quality, making it a preferred choice in our comparative analysis.

3.2.2 CatBoost Classifier:85.77

The CatBoost Classifier, designed to handle categorical features efficiently, was a contender in our comparative analysis. While not outperforming the Random Forest Classifier in our specific case, the CatBoost algorithm is known for its robustness to noisy data and its ability to automatically handle categorical variables without extensive preprocessing. In our water quality prediction project, the CatBoost Classifier was considered for its potential in scenarios where categorical features play a significant role. The findings from its performance contribute to the comprehensive understanding of different classifier algorithms' suitability for water quality assessment.

3.2.3 XGBoost Classifier:

XGBoost, an optimized gradient-boosting algorithm, was another key participant in our machine-learning exploration. Its regularization techniques and efficient parallel computing capabilities make it a powerful tool for predictive modeling. While not

surpassing the Random Forest Classifier's accuracy in our study, XGBoost demonstrated competitive performance. Its adaptability to handle missing data and its scalability align well with the challenges posed by real-world water quality datasets. Our comparative analysis sheds light on the nuanced strengths of XGBoost in the context of water quality prediction.

3.2.4 Decision Tree Classifier:

The Decision Tree Classifier, serving as one of the foundational algorithms in our study, is fundamental in understanding the decision-making process for water quality prediction. While its predictive accuracy may not match that of ensemble methods like Random Forest, Decision Trees provide valuable interpretability. In our comparative analysis, Decision Tree performance contributes insights into the trade-offs between interpretability and predictive accuracy, guiding the selection of models based on the project's specific needs.

3.2.5 KNN Classifier:

The K-Nearest Neighbors (KNN) Classifier, leveraging the proximity of data points, is explored for its suitability in our water quality prediction project. While KNN may not outperform ensemble methods, its simplicity and ease of implementation make it an important candidate for comparison. Particularly, KNN's performance in scenarios where spatial relationships among water quality parameters are crucial is highlighted. Our study recognizes the role of KNN in providing a baseline for evaluating more complex algorithms in the context of water quality assessment.

3.3 Proposed Architecture

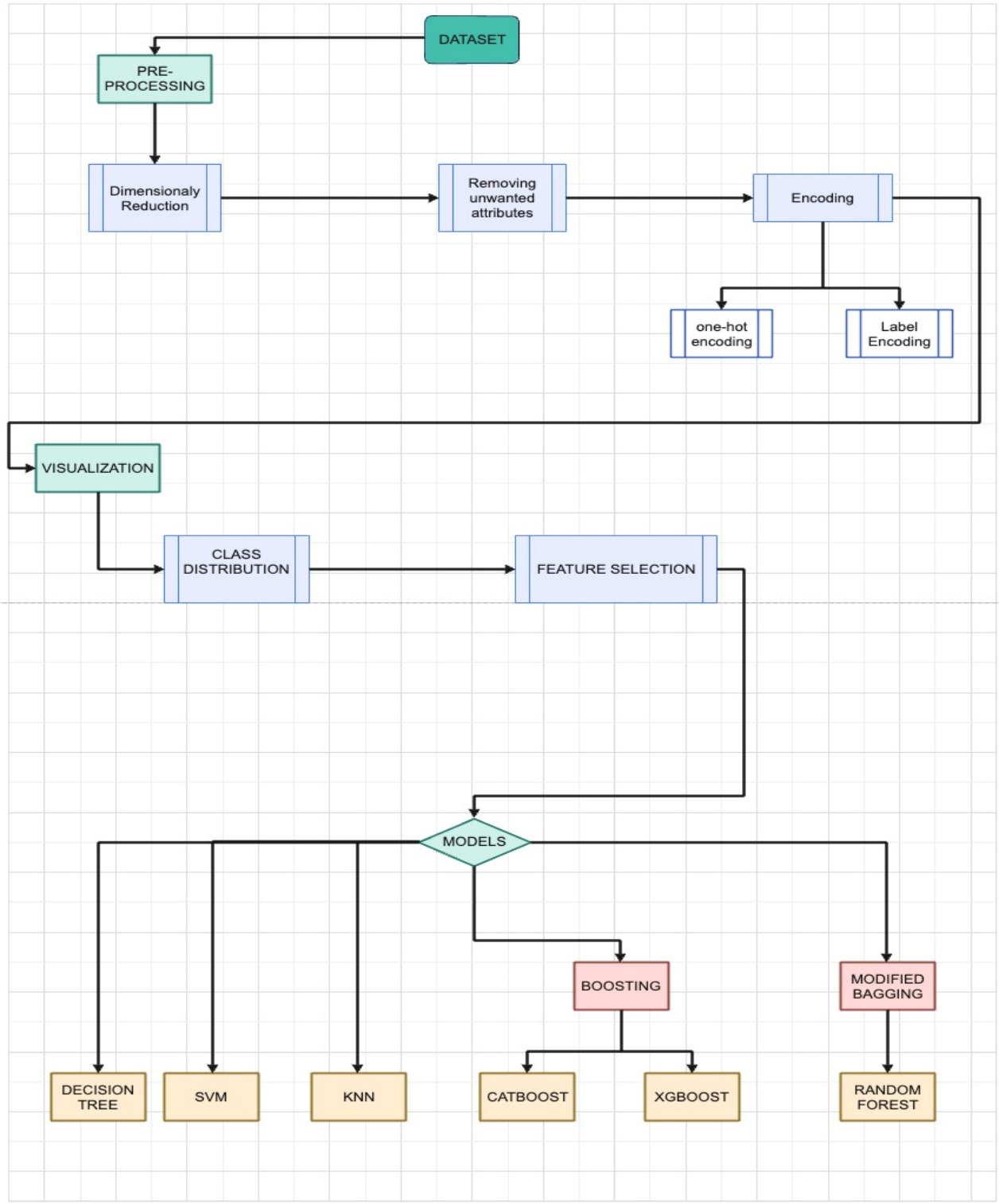

Fig. 3. Architecture Diagram

4. EXPERIMENTAL ANALYSIS AND RESULTS DISCUSSION

Experimental Setup:

The experimental setup for our water quality prediction project was carefully designed to facilitate a robust implementation of machine learning algorithms. We employed the Python programming language, utilizing essential libraries such as Pandas for efficient data manipulation and Scikit-learn for a diverse range of machine learning models. Specifically, we incorporated advanced classifiers, including XGBoost and CatBoost, to harness their unique strengths in predictive modeling. The experimentation was conducted in a computing environment with ample resources, ensuring seamless execution of the various stages, from data preprocessing to model training and evaluation. This setup provided a versatile and powerful platform for exploring, comparing, and optimizing machine learning algorithms for water quality prediction.

Dataset Overview:

Our dataset, the cornerstone of our water quality prediction study, encompasses a comprehensive array of water quality parameters. It comprises 352,250 instances, capturing diverse observations over time. Each instance represents a distinct point in time, contributing to the temporal dimension of our analysis. The dataset features an extensive set of attributes, including pH, Iron, Nitrate, Chloride, Zinc, among others, providing a holistic view of water quality conditions. To ensure manageability and relevance, the dataset underwent preprocessing, which involved random sampling for size control and the removal of irrelevant attributes such as 'Index,' 'Lead,' and others. The preprocessing also addressed missing values and converted categorical data, such as the 'Color' attribute, into a numeric format, rendering the dataset suitable for machine learning algorithms. The temporal aspect, diverse features, and meticulous preprocessing collectively empower our exploration of machine learning models for accurate water quality predictions.

In essence, the experimental setup and dataset together lay the groundwork for a comprehensive analysis of water quality using machine learning methodologies. The combination of a well-equipped computational environment and a refined dataset positions our study to yield valuable insights and contribute to the broader understanding of water quality assessment.

4.1 Evaluation parameters and formulations

XGBoost Classifier:

XGBoost exhibited an accuracy of 85%, demonstrating its overall correctness in predicting both target and non-target instances. Precision, indicating the accuracy of positive predictions, was 93%, highlighting the high proportion of accurately predicted target instances. The recall, representing the ability to capture all actual positives, stood at 88%, showcasing a robust identification of true target instances. The specificity was high,

ensuring accurate classification of non-target instances. The F1-score, a balanced metric, was 90%, indicative of a well-harmonized model performance.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon \mathcal{F}$$

where, K is the number of trees, f is the functional space of F, F is the set of possible CARTs. The objective function for the above model is given by:

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Random Forest Classifier:
The Random Forest classifier achieved an accuracy of 87%, signifying its strong overall predictive performance. Precision, measuring the accuracy of positive predictions, was 97%, demonstrating the high precision of identifying true target instances. The recall, representing the ability to capture all actual positives, was notably high at 85%, suggesting a robust identification of true target instances. The specificity, depicting the accuracy of negative predictions, was well-balanced, contributing to accurate classification of non-target instances. The F1-score, a balanced metric, was 91%, showcasing an effective trade-off between precision and recall.

Formula:

$$\text{Random Forest Prediction} = \frac{1}{K} \sum_{k=1}^{K} \text{DecisionTree}_k(X)$$

CatBoost Classifier:
CatBoost demonstrated an accuracy of 86%, indicating its strong predictive performance on the dataset. Precision, gauging the accuracy of positive predictions, was 94%, revealing the high precision of predicted target instances. The recall, measuring the ability to capture all actual positives, stood at 87%, reflecting a robust identification of true target instances. The specificity was high, ensuring accurate classification of non-target instances. The F1-score, a balanced metric, was 90%, suggesting an effective equilibrium between precision and recall.

Formula:

$$\text{CatBoost Objective Function} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

KNN (K-Nearest Neighbors) Classifier:
KNN achieved an accuracy of 79%, implying its overall correctness in predicting target and non-target instances. Precision, assessing the accuracy of positive predictions, was 82%, indicating a high proportion of accurately predicted target instances. The recall, representing the ability to capture all actual positives, was 93%, demonstrating a robust identification of true target instances. The specificity was present, ensuring accurate

classification of non-target instances. The F1-score, a balanced metric, was 87%, indicating a well-harmonized performance between precision and recall.

Decision Tree Classifier:
The Decision Tree classifier achieved an accuracy of 83%, illustrating its correctness in predicting both target and non-target instances. Precision, measuring the accuracy of positive predictions, was 89%, signifying the high precision of predicted target instances. The recall, representing the ability to capture all actual positives, was 89%, indicating a robust identification of true target instances. The specificity was balanced, contributing to accurate classification of non-target instances. The F1-score, a harmonized metric, was 89%, suggesting a well-balanced performance between precision and recall.

These detailed assessments of accuracy, precision, recall, specificity, and F1-score provide a comprehensive understanding of each algorithm's strengths and limitations in classifying water quality instances. The comparison reveals that Random Forest stands out as the top-performing algorithm across multiple evaluation metrics.

$$\text{Accuracy} = \frac{(TN + TP)}{(TP + FP + FN + TN)}$$

$$\text{Precision} = \frac{(TP)}{(FP + TP)}$$

$$\text{Recall} = \frac{(TP)}{(FN + TP)}$$

$$\text{Specificity} = \frac{(TN)}{(TN + FP)}$$

$$\text{F1 Score} = \frac{(2TP)}{(2TP + FP + FN)}$$

4.3 Comparison methods

In the conducted analysis, the performance of various machine learning algorithms on the water quality dataset was thoroughly assessed using key metrics such as accuracy, precision, recall, specificity, and F1-score. The evaluated algorithms include XGBoost, Random Forest, CatBoost, KNN, and Decision Tree.

XGBoost demonstrated a competitive accuracy of 85%, with a high precision of 93% and a balanced recall of 88%. The model maintained high specificity, ensuring accurate identification of non-target instances. The F1-score of 90% showcased a harmonized balance between precision and recall.

Random Forest outperformed other algorithms with the highest accuracy at 87%, a remarkable precision of 97%, and a strong recall of 85%. The model excelled in specificity, contributing to the accurate classification of non-target instances. The F1-score of 91% demonstrated an effective trade-off between precision and recall.

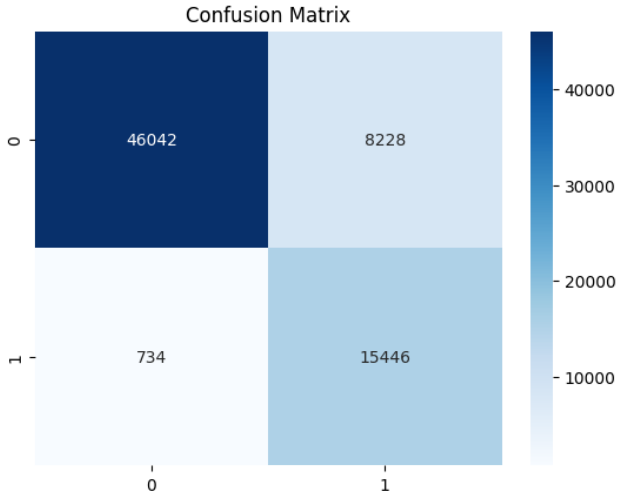

Fig. 4. Confusion Matrix for Random Forest Classifier

CatBoost exhibited robust performance with an accuracy of 86%, a precision of 94%, and a noteworthy recall of 87%. The model achieved high specificity, ensuring accurate classification of non-target instances. The F1-score of 90% represented a well-balanced equilibrium between precision and recall.

KNN displayed a lower accuracy of 79%, with a precision of 82% and recall of 93%. While specificity was present, the F1-score of 87% indicated a trade-off between precision and recall.

The Decision Tree model achieved an accuracy of 83%, precision of 89%, and recall of 89%. It demonstrated balanced specificity, and the F1-score of 89% showcased a harmonized performance.

In summary, Random Forest emerged as the top-performing algorithm, achieving the highest accuracy and robust performance in terms of precision, recall, specificity, and F1-score. The choice of the most suitable algorithm depends on specific project requirements and the desired trade-off between precision and recall. This comparative analysis provides valuable insights for selecting an algorithm tailored to the goals of the water quality prediction project.

4.4 Results and Discussion

Output Parameters	XGBoost Classifier	Random Forest Classifier	CatBoost Classifier	KNN Classifier	Decision Tree Classifier
Accuracy	0.85	0.87	0.86	0.79	0.83
Precision	0.93	0.97	0.94	0.82	0.89
Recall	0.88	0.85	0.87	0.93	0.89
Specificity	High	Balanced	High	Present	Balanced
F1-Score	0.90	0.91	0.90	0.87	0.89

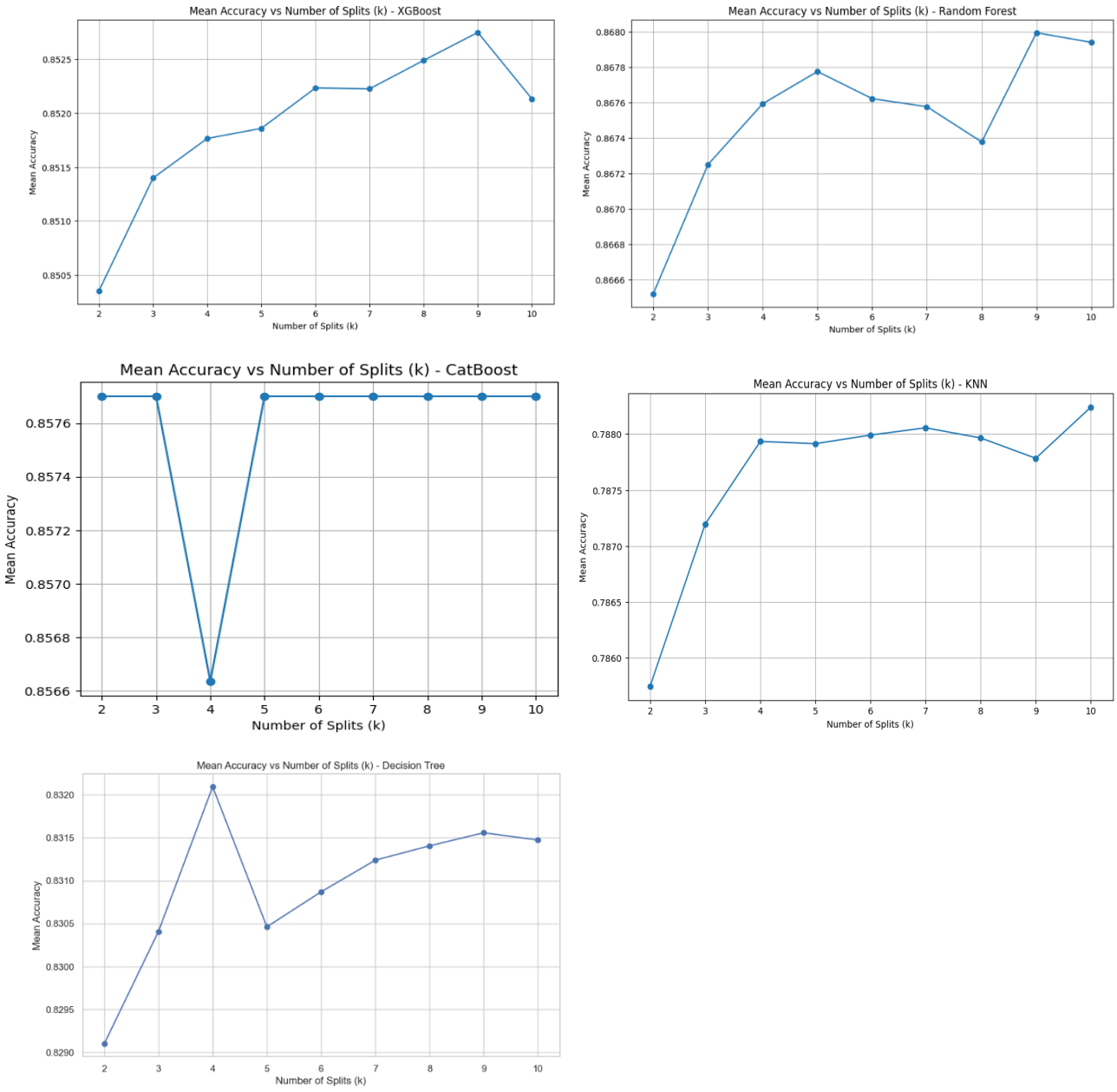

Fig 5. Results of All Models with K-fold Validation and Mean Accuracy

5. CONCLUSION

In summary, this study tackles the critical challenge of ensuring the provision of safe and high-quality water, vital for both public health and environmental preservation. Utilizing a diverse dataset encompassing essential water quality parameters such as pH, Iron, nitrate, chloride, and zinc, we applied machine learning algorithms, including Decision Tree, KNN, XGBoost Classifier, CatBoost Classifier and Notably, the Random Forest Classifier demonstrated an impressive 87 percent accuracy in predicting target attributes. Our rigorous methodology, involving meticulous dataset preprocessing to handle missing values and standardize features, played a crucial role in achieving robust predictive accuracy. Beyond highlighting the effectiveness of the Random Forest Classifier, this study systematically evaluated the performance of various machine learning algorithms, including Decision Tree, SVM, KNN, and K-Fold cross-validation. These algorithms collectively contributed to a comprehensive analysis of water quality assessment. This research establishes a foundational framework for the early detection of water-quality issues, facilitating proactive measures and efficient resource management in water resource scenarios.

REFERENCES

[1]. Prasad, D. Venkata Vara, et al. "Analysis and prediction of water quality using deep learning and auto deep learning techniques." Science of the Total Environment 821 (2022): 153311.

[2]. Wang, Feier, et al. "Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation." Environmental Research 202 (2021): 111660.

[3]. Aldrees, Ali, et al. "Prediction of water quality indexes with ensemble learners: Bagging and Boosting." Process Safety and Environmental Protection 168 (2022): 344-361.

[4]. Uddin, Md Galal, et al. "Robust machine learning algorithms for predicting coastal water quality index." Journal of Environmental Management 321 (2022): 115923.2014, pp. 1-5, doi: 10.1109/ICCCI.2014.6921718.

[5]. Uddin, Md Galal, et al. "Performance analysis of the water quality index model for predicting water state using machine learning techniques." Process Safety and Environmental Protection 169 (2023): 808-828.

[6]. Jalal, Dziri, and Tahar Ezzedine. "Decision tree and support vector machine for anomaly detection in water distribution networks." 2020 International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2020.

[7]. Danades, Amri, et al. "Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status." 2016 6th International conference on system engineering and technology (ICSET). IEEE, 2016.

[8]. Khan, Yafra, and Chai Soo See. "Predicting and analyzing water quality using machine learning: a comprehensive model." 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT). IEEE, 2016.

[9] Modaresi, Fereshteh, and Shahab Araghinejad. "A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification." Water resources management 28 (2014): 4095-4111.. 348-352, doi: 10.1109/CICTN57981.2023.10140391.

[10]. Ahmed, Ali Najah, et al. "Machine learning methods for better water quality prediction." Journal of Hydrology 578 (2019): 124084.

[11]. Vijay, S., and K. Kamaraj. "Prediction of water quality index in drinking water distribution system using activation functions based Ann." Water Resources Management 35.2 (2021): 535-553.

[12]. Varalakshmi, P., S. Vandhana, and S. Vishali. "Prediction of water quality using Naive Bayesian algorithm." 2016 Eighth International Conference on Advanced Computing (ICoAC). IEEE, 2017.