IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

WATER QUALITY PREDICTION USING MACHINE LEARNING

K.TULASI KRISHNA KUMAR, G.NOSHNA

Assistant Professor & Training & Placement Officer, 2 MCA Final Semester,
Master of Computer Applications,
Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India.

Abstract: Water is the most crucial resource of life and it is necessary for the survival of all living creatures including human beings. The survival of business and agriculture depends on fresh water. An essential step in managing freshwater assets is the evaluation of the quality of the water. Before using water for anything, including drinking, chemical spraying (pesticides, etc.),or animal hydration , it is crucial to assess its purity. The ecosystem and the general public's health are directly impacted by water quality. Therefore, analyzing and predicting water quality is necessary for both environment and human protection. Machine learning can be used to analyze and predict the water quality based on the parameters like PH value, turbidity, hardness, conductivity, dissolved solids in water and others parameters as input to machine learning algorithms and the water is classified as safe or unsafe for the usage of domestic purposes. The Flask applications predicts water quality safety using an XG Boost classifier trained on a dataset (waterQuality1.csv). The dataset undergoes preprocessing where missing values are dropped and object type columns are converted to numeric types. Categorical target variables are encoded for machine learning compatibility. The model is trained on the processed data, evaluated for accuracy, and then used to predict the safest status ("Safe" or "Not Safe")based on user-inputted water quality parameters. Predicted results are displayed ton users on a web page rendered using Flask (predictdailyhousehold.html).The proposed work uses various ML models such as Logistic Regression, Support Vector Machine (SVM), XG Boost(XG) and Random Forest(RF) to classify whether the water is drinkable. The XG boost classifier is selected for the explanation and yields optimum Accuracy and F1-Score of 0.98, with Precision and Re-call of 0.97 and 0.99 respectively. This work is an emerging research at present with a vision of addressing the water quality for the future as well.

Index Terms - XG Boost classifier, Support Vector Machine (SVM), Logistic Regression, Random Forest, Robust Scaler, Exploratory Data Analysis

I. INTRODUCTION

Access to clean and safe water is a fundamental human necessity, yet water contamination remains a pressing global issue, particularly in developing nations where environmental regulations and water treatment infrastructure may be inadequate. Poor water quality poses severe health risks and environmental hazards, including the spread of waterborne diseases and degradation of ecosystems. Monitoring and ensuring the safety of water resources is thus a key priority for governments, environmental agencies, and public health organizations [5]. Traditionally, water quality analysis has relied heavily on manual sampling and laboratory testing, which are time-consuming, labor-intensive, and not always feasible for real-time monitoring. With the advancement of data-driven technologies, there has been a growing shift toward using machine learning techniques to automate and enhance water quality assessment. These techniques offer the capability detect patterns and predict water safety levels based on various chemical, physical, and biological indicators present in the water. In this project, we apply several popular machine learning algorithms—namely Random Forest, Support Vector Machine (SVM), Logistic Regression, and XG Boost—to classify water as "safe" or "unsafe" based on a variety of quality parameters. The dataset, once preprocessed and standardized, was used to train and evaluate these models, with the goal of identifying the most effective technique for accurate water safety classification. Standard metrics such as accuracy, confusion matrix, and classification reports were used to assess the performance of each model [1]. Water quality prediction, like many classification problems, comes with its own challenges, such as class imbalance, noisy data, and the need for appropriate feature selection. Therefore, significant emphasis was placed on data preprocessing, label encoding, and feature scaling to ensure optimal model performance [14]. Ultimately, our findings aim to demonstrate the feasibility of using machine learning for environmental monitoring and contribute to the development of intelligent systems for public health and resource management**.**

2. Literature Survey

Water quality prediction has become an increasingly important area of research due to its implications for public health, environmental protection, and sustainable resource management. A wide range of techniques have been proposed over the years, with particular emphasis on the use of neural networks, data mining, and distributed data mining methods. These techniques aim to model and predict water quality based on various physicochemical and biological parameters collected from rivers, lakes, groundwater, or municipal supplies. Numerous studies have demonstrated the effectiveness of machine learning and related methods in this domain. The most commonly used algorithms include artificial neural networks (ANN), rule-induction techniques, decision trees, logistic regression, and support vector machines (SVM) [1]. These models are capable of identifying complex, non-linear relationships within the data, making them well-suited for handling the variability and dynamic nature of water quality indicators. In their study, Behrouz, Emad, and Sahil [2] proposed a water quality classification approach using supervised machine learning algorithms. They implemented an ensemble learning strategy to create a "super classifier" that combines the strengths of multiple models. Their work not only demonstrated improved prediction accuracy but also provided a comparative analysis of traditional supervised learning algorithms against their ensemble-based approach. Similarly, Navanshu Khare and Saad Yunus Sait [3] offered in-depth insights into the theoretical foundations of various machine learning algorithms employed for water quality prediction. Their study emphasized understanding the mathematical operations and logic behind the algorithms to better interpret model results and performance. Further, Siddhartha Bhattacharyya and colleagues [6], in their 2011 study, conducted a detailed comparative analysis of Support Vector Machines, Random Forests, and Logistic Regression models in the context of water quality classification. Through extensive experiments, they concluded that the Random Forest algorithm provided the highest classification accuracy, followed by Logistic Regression and Support Vector Machines, particularly in handling complex environmental datasets. In another noteworthy study, Y. Sahin and E. Duman [7] explored a hybrid modeling approach using both Support Vector Machines and Decision Trees. Their research highlighted that while Decision Trees offered better performance on smaller datasets due to their simplicity and interpretability, Support Vector Machines outperformed as the dataset size increased, owing to their strong generalization capabilities.

3.Challenges

Despite the advantages, there are several challenges in building effective water quality prediction systems:

- **Data Availability:** Water quality datasets are not always readily available and often lack consistency or completeness.
- **Regional Differences:** Water properties can vary significantly across different regions and climates, affecting model generalizability.
- **Imbalanced Data:** In many datasets, the number of unsafe samples is much smaller than safe ones, leading to imbalanced class distributions.
- **Model Interpretability**: While complex models like XG Boost are powerful, they are less interpretable, making it difficult to explain predictions to non-technical stakeholders.to explain predictions to non-technical stakeholders

4. Exploratory Data Analysis and Other Findings

As noted earlier, the dataset is highly imbalanced, which is evident from the bar plot below. Only a small fraction of the water samples Are classified as poor quality. This indicates that the data is significantly unbalanced with respect to the target variable Water Quality Prediction. We can see that there is no significant correlation among the independent variables, which is expected as the data providers have applied Principal Component Analysis (PCA) to the input features, except for variables such as Sampling Time and Contaminant Concentration. There are some minor correlations between certain features and Sampling Time (e.g., an inverse correlation with PC3) and Contaminant Concentration (e.g., a positive correlation with PC7 and PC20, and an inverse correlation with PC1 and PC5).

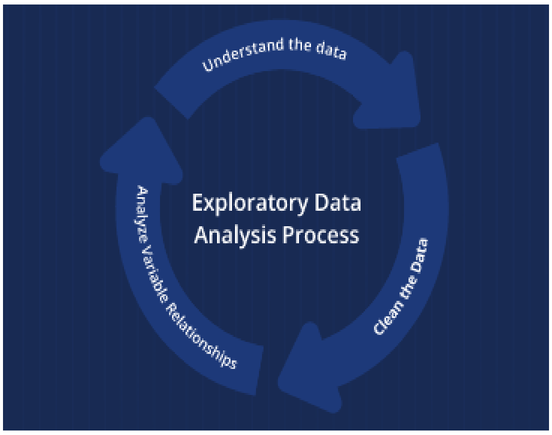

Understand the data
Clean the Data
Analyze Variable Relationships
Exploratory Data Analysis Process

Figure 1 Exploratory Data Analysis

5. SYSTEM DESIGN

Functional Requirements

Functional requirements define the core capabilities that the system must offer to fulfill its intended purpose. These requirements focus on what the system should do.

Non-Functional Requirements

Non-functional requirements specify how the system performs its functions and define quality attributes such as reliability, usability, and performance.

- **Availability**: The system should be available for use at all times, ensuring accessibility for environmental agencies, research institutions, or individual users. It should support both online and offline usage (in a local setup).
- **Accuracy and Correctness**: Prediction accuracy is a key requirement for a critical system like this. The model should be trained on validated datasets to minimize errors. The accuracy level should be as high as possible, ideally exceeding 90%, to provide users with confidence in the system's results.
- **Maintainability**: The system should be designed in a modular way so that it can be easily maintained and updated. It should support future modifications such as adding new features, updating datasets, or integrating advanced machine learning models without affecting the core functionality.
- **Usability**: The interface should be intuitive and user-friendly, ensuring that users without technical expertise can easily interact with the system, input data, and understand the predictions. Clear labels, help documentation, and error-handling messages should be provided.
- **Scalability and Flexibility**: The system should be scalable to support larger datasets in the future and allow integration with real-time data sources like IoT sensors or water monitoring devices.

6.Dataset

The dataset used for this project contains measurements of various water quality parameters collected from multiple sources, such as rivers, lakes, or municipal water supplies. Each record in the dataset includes features like:

- **pH** – acidity or alkalinity of water
- **Turbidity** – clarity of water
- **Dissolved Oxygen (DO)** – oxygen level required for aquatic life
- **Biological Oxygen Demand (BOD)** – the amount of oxygen consumed by microorganisms
- **Chemical Oxygen Demand (COD)** – total measurement of all chemicals in the water
- **Total Dissolved Solids (TDS)** – concentration of dissolved substances
- **Temperature**, **Conductivity**, and more

The target variable is typically either a **Water Quality Index (WQI)** score (continuous value) or a **categorical label** (e.g., "Safe" vs. "Unsafe").

The dataset may include:

- Historical records collected from lab measurements or IoT sensors
- Both labeled and unlabeled data
- Multiple classes or binary classification problems

Most datasets are **cleaned and preprocessed** before modeling. This includes normalization, handling missing values, and feature selection. In many cases, **no null values** are present, simplifying the preprocessing phase.

Models are trained using supervised learning techniques on a labeled dataset (with known water quality outcomes) and validated using techniques like **train-test split**, **cross-validation**, and **performance metrics** such as accuracy, F1-score, and confusion matrix.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

The water quality dataset is highly imbalanced, with a large portion of the records labeled as safe and only a small fraction labeled as unsafe. This imbalance means that most of the water samples in the dataset are considered safe for use. If we use this dataset directly as the base for training our predictive models, we may achieve high accuracy scores without effectively identifying unsafe water samples. This happens because the model might "assume" that most water samples are safe, leading to biased predictions. However, our goal is not to assume safety but to enable the model to learn meaningful patterns that can accurately distinguish between safe and unsafe water. Therefore, handling class imbalance is crucial for building a reliable and robust water quality prediction model

1. Dataset Input
2. Dataset Analysis
3. Oversampling (using SMOTE)
4. Training and Testing Subset
5. Using the algorithm
7.Making Predictions about Outcomes

6.1 Dataset Input:

You can get the dataset from an online data provider using online resources. In order to accurately estimate the accuracy, we must amass a sizable collection of data.

6.2 Dataset analysis:

This section contains dataset analysis. For the data processing, the data size is taken into account.

6.3 Directly applying the classification:

The results reveal a challenge known as the accuracy paradox. The accuracy paradox refers to the misleading nature of accuracy as a performance metric when dealing with imbalanced datasets, such as in the case of water quality prediction [9]. In our dataset, the majority of samples are labeled as safe, which can lead a predictive model to achieve high accuracy simply by predicting every instance as safe. However, such a model fails to identify unsafe water samples, rendering it practically ineffective. For example, if 99% of the samples are safe, a model predicting "safe" for every input will have 99% accuracy—but it completely misses the point of detecting unsafe conditions. In such cases, metrics like precision, recall, and F1-score are more reliable, as they provide insights into how well the model distinguishes between safe and unsafe water. The root of this issue lies in the class imbalance between the safe and unsafe categories. Hence, it is important to consider the prior probabilities of these classes during error analysis. While precision and recall help evaluate model performance more appropriately, even precision can be affected by extreme imbalance in class distributions. Furthermore, by examining feature distributions, we can observe how skewed some of the features are[11]. In future work, techniques to reduce skewness and balance class distribution—such as oversampling, under sampling, or data transformation—will be applied to improve model robustness and fairness. To handle the issue of class imbalance in water quality prediction, an effective approach is to adjust the class distribution by sampling the minority class. In our case, the minority class represents unsafe water samples. By increasing the number of training examples for the unsafe class, we can provide the model with more opportunities to learn meaningful patterns associated with unsafe conditions. This process, known as oversampling, helps balance the dataset and improve the model's ability to correctly identify unsafe water. By training the algorithm with a more representative sample of both classes, the likelihood of accurate and fair predictions increases, reducing the bias toward the majority (safe) class. This technique plays a crucial role in ensuring that the model performs well not just overall, but specifically in detecting unsafe water—where correct classification is most critica

7.Scaling and Sub-sampling:

In this phase, we focus on scaling the numerical features and balancing the dataset to enhance model performance. Just like other features, certain columns in the water quality dataset such as measurement-related features (e.g., pH, turbidity, dissolved oxygen)—need to be standardized to ensure they are on a comparable scale. This step is essential for models like SVM and Logistic Regression, which are sensitive to feature scaling. Additionally, due to the significant class imbalance in the original dataset (with many more safe water samples than unsafe ones), we create a sub sample to achieve a balanced representation of both classes. In this scenario, the sub sample will have a 50/50 ratio of safe and unsafe water samples, allowing the machine learning algorithms to better learn patterns associated with water that is unsafe for use. This balanced sub sample is necessary because the original dataset can lead to:

Over fitting: Classification models may learn to assume most samples are safe, failing to detect unsafe water. We want our models to be sensitive and accurate when identifying unsafe conditions.

Misleading Correlations: With imbalanced data, the relationships between features and the target class (safe or unsafe) can be obscured. By balancing the data, we can more accurately analyze the influence of each feature on water safety classification and build more reliable models

In our water quality prediction system, we utilize Robust Scaler to scale the input features. This is particularly beneficial because Robust Scaler scales features using statistics that are resistant to outliers, which are commonly present in environmental and sensor-based datasets like water quality measurements. The Robust Scaler works by removing the median and scaling the data according to the interquartile range (IQR)—the range between the 25th percentile (Q1) and the 75th percentile (Q3). This ensures that the transformation is less influenced by extreme values, unlike traditional methods such as standard scaling (which relies on mean and standard deviation and is highly sensitive to outliers). During training, centering and scaling are done independently for each feature using the median and IQR calculated from the training set. These statistics are then stored and used to transform any future data consistently. This approach ensures standardization without being skewed by anomalous readings, [13] which is crucial for building stable and generalizable machine learning models. In comparison, methods like Quantile Transformer are also robust to outliers and additionally transform the data into a uniform or normal distribution. Unlike Robust Scaler, Quantile Transformer collapses outliers to predefined boundaries (e.g., 0 and 1), which can be beneficial in some applications but may distort feature relationships in others.

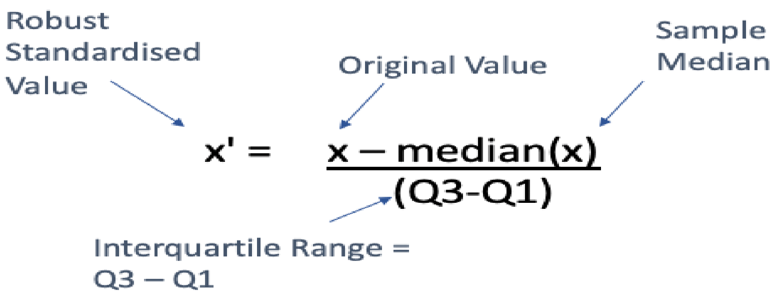

Robust Standardised Value
Original Value
Sample Median
x' = x − median(x) / (Q3-Q1)
Interquartile Range = Q3 − Q1

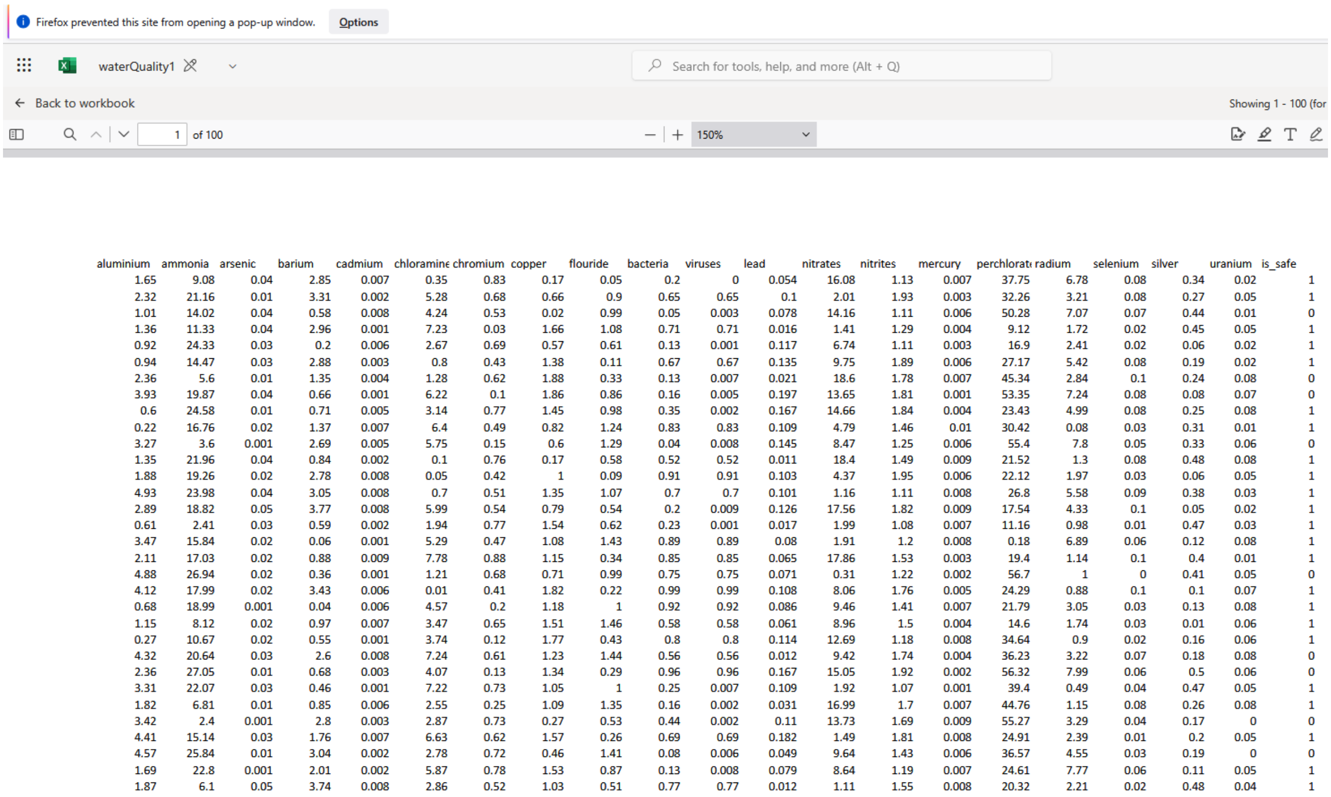
Figure: 2 – Robust Sclar Method

8. Result Analysis

In this phase, we focus on scaling the numerical features and balancing the dataset to enhance model performance. Just like other features, certain columns in the water quality dataset—such as measurement-related features (e.g., pH, turbidity, dissolved oxygen)—need to be standardized to ensure they are on a comparable scale. This step is essential for models like SVM and Logistic Regression, which are sensitive to feature scaling. Additionally, due to the significant class imbalance in the original dataset (with many more safe water samples than unsafe ones), we create a subsample to achieve a balanced representation of both classes. In this scenario, the subsample will have a 50/50 ratio of safe and unsafe water samples, allowing the machine learning algorithms to better learn patterns associated with water that is unsafe for use. This balanced subsample is necessary because the original dataset can lead to:

- **Overfitting:** Classification models may learn to assume most samples are safe, failing to detect unsafe water. We want our models to be sensitive and accurate when identifying unsafe conditions.
- **Misleading Correlations:** With imbalanced data, the relationships between features and the target class (safe or unsafe) can be obscured. By balancing the data, we can more accurately analyze the influence of each feature on water safety classification and build more reliable models.

8.1 INPUT SCREENS:

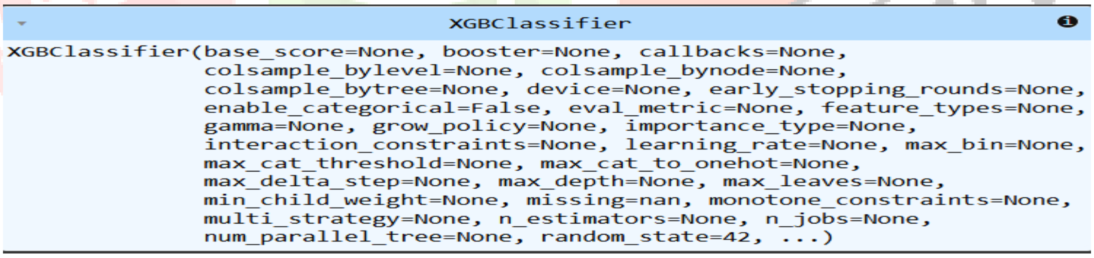


Firefox prevented this site from opening a pop-up window.

aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride	bacteria	viruses	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	is_safe
1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.2	0	0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1
2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.9	0.65	0.65	0.1	2.01	1.93	0.003	32.26	3.21	0.08	0.27	0.05	1
1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	0.003	0.078	14.16	1.11	0.006	50.28	7.07	0.07	0.44	0.01	0
1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	0.71	0.016	1.41	1.29	0.004	9.12	1.72	0.02	0.45	0.05	1
0.92	24.33	0.03	0.2	0.006	2.67	0.69	0.57	0.61	0.13	0.001	0.117	6.74	1.11	0.003	16.9	2.41	0.02	0.06	0.02	1
0.94	14.47	0.03	2.88	0.003	0.8	0.43	1.38	0.11	0.67	0.67	0.135	9.75	1.89	0.006	27.17	5.42	0.08	0.19	0.02	1
2.36	5.6	0.01	1.35	0.004	1.28	0.62	1.88	0.33	0.13	0.007	0.021	18.6	1.78	0.007	45.34	2.84	0.1	0.24	0.08	0
3.93	19.87	0.04	0.66	0.001	6.22	0.1	1.86	0.86	0.16	0.005	0.197	13.65	1.81	0.001	53.35	7.24	0.08	0.08	0.07	0
0.6	24.58	0.01	0.71	0.005	3.14	0.77	1.45	0.98	0.35	0.002	0.167	14.66	1.84	0.004	23.43	4.99	0.08	0.25	0.08	1
0.22	16.76	0.02	1.37	0.007	6.4	0.49	0.82	1.24	0.83	0.83	0.109	4.79	1.46	0.01	30.42	0.08	0.03	0.31	0.01	1
3.27	3.6	0.001	2.69	0.005	5.75	0.15	0.6	1.29	0.04	0.008	0.145	8.47	1.25	0.006	55.4	7.8	0.05	0.33	0.06	0
1.35	21.96	0.04	0.84	0.002	0.1	0.76	0.17	0.58	0.52	0.52	0.011	18.4	1.49	0.009	21.52	1.3	0.08	0.48	0.08	1
1.88	19.26	0.02	2.78	0.008	0.05	0.42	1	0.09	0.91	0.91	0.103	4.37	1.95	0.006	22.12	1.97	0.03	0.06	0.05	1
4.93	23.98	0.04	3.05	0.008	0.7	0.51	1.35	1.07	0.7	0.7	0.101	1.16	1.11	0.008	26.8	5.58	0.09	0.38	0.03	1
2.89	18.82	0.05	3.77	0.008	5.99	0.54	0.79	0.54	0.2	0.009	0.126	17.56	1.82	0.009	17.54	4.33	0.1	0.05	0.02	1
0.61	2.41	0.03	0.59	0.002	1.94	0.77	1.54	0.62	0.23	0.001	0.017	1.99	1.08	0.007	11.16	0.98	0.01	0.47	0.03	1
3.47	15.84	0.02	0.06	0.001	5.29	0.47	1.08	1.43	0.89	0.89	0.08	1.91	1.2	0.008	0.18	6.89	0.06	0.12	0.08	1
2.11	17.03	0.02	0.88	0.009	7.78	0.88	1.15	0.34	0.85	0.85	0.065	17.86	1.53	0.003	19.4	1.14	0.1	0.4	0.01	1
4.88	26.94	0.02	0.36	0.001	1.21	0.68	0.71	0.99	0.75	0.75	0.071	0.31	1.22	0.002	56.7	1	0	0.41	0.05	0
4.12	17.99	0.02	3.43	0.006	0.01	0.41	1.82	0.22	0.99	0.99	0.108	8.06	1.76	0.005	24.29	0.88	0.1	0.1	0.07	1
0.68	18.99	0.001	0.04	0.006	4.57	0.2	1.18	1	0.92	0.92	0.086	9.46	1.41	0.007	21.79	3.05	0.03	0.13	0.08	1
1.15	8.12	0.02	0.97	0.007	3.47	0.65	1.51	1.46	0.58	0.58	0.061	8.96	1.5	0.004	14.6	1.74	0.03	0.01	0.06	1
0.27	10.67	0.02	0.55	0.001	3.74	0.12	1.77	0.43	0.8	0.8	0.114	12.69	1.18	0.008	34.64	0.9	0.02	0.16	0.06	1
4.32	20.64	0.03	2.6	0.008	7.24	0.61	1.23	1.44	0.56	0.56	0.012	9.42	1.74	0.004	36.23	3.22	0.07	0.18	0.08	0
2.36	27.05	0.01	0.68	0.003	4.07	0.13	1.34	0.29	0.96	0.96	0.167	15.05	1.92	0.002	56.32	7.99	0.06	0.5	0.06	0
3.31	22.07	0.03	0.46	0.001	7.22	0.73	1.05	1	0.25	0.007	0.109	1.92	1.07	0.001	39.4	0.49	0.04	0.47	0.05	1
1.82	6.81	0.01	0.85	0.006	2.55	0.25	1.09	1.35	0.16	0.002	0.031	16.99	1.7	0.007	44.76	1.15	0.08	0.26	0.08	1
3.42	2.4	0.001	2.8	0.003	2.87	0.73	0.27	0.53	0.44	0.002	0.11	13.73	1.69	0.009	55.27	3.29	0.04	0.17	0	0
4.41	15.14	0.03	1.76	0.007	6.63	0.62	1.57	0.26	0.69	0.69	0.182	1.49	1.81	0.008	24.91	2.39	0.01	0.2	0.05	1
4.57	25.84	0.01	3.04	0.002	2.78	0.72	0.46	1.41	0.08	0.006	0.049	9.64	1.43	0.006	36.57	4.55	0.03	0.19	0	0
1.69	22.8	0.001	2.01	0.002	5.87	0.78	1.53	0.87	0.13	0.008	0.079	8.64	1.19	0.007	24.61	7.77	0.06	0.11	0.05	1
1.87	6.1	0.05	3.74	0.008	2.86	0.52	1.03	0.51	0.77	0.77	0.012	1.11	1.55	0.008	20.32	2.21	0.02	0.48	0.04	1

Figure 3: CSV file as input

8.2 OUTPUT SCREENS:



```
XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=None, n_jobs=None,
              num_parallel_tree=None, random_state=42, ...)
```

Figure 4 - XGB Classifier

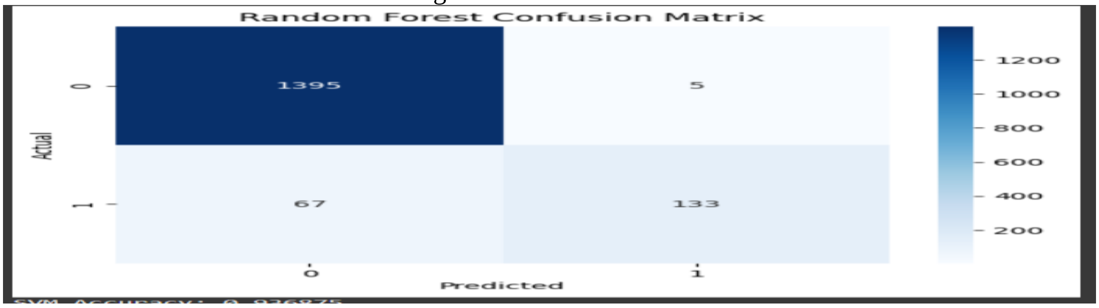


Figure 5 - Random Forest Confusion Matrix

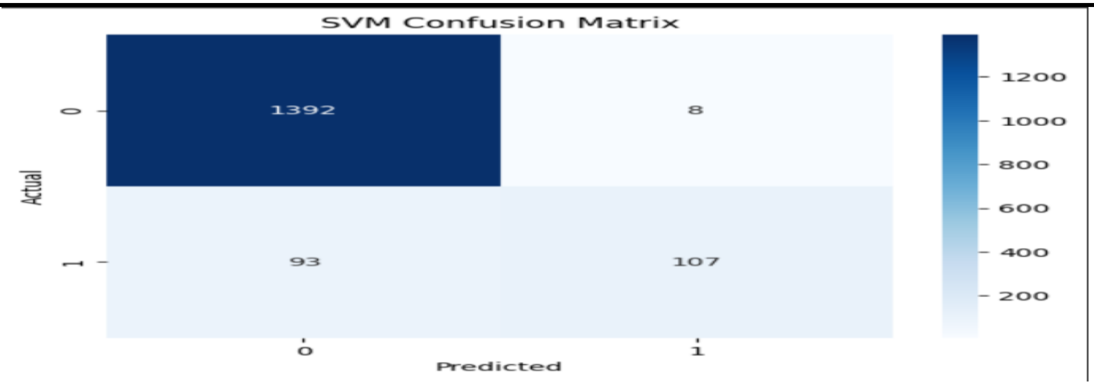


Fig 6- SVM Confusion Matrix

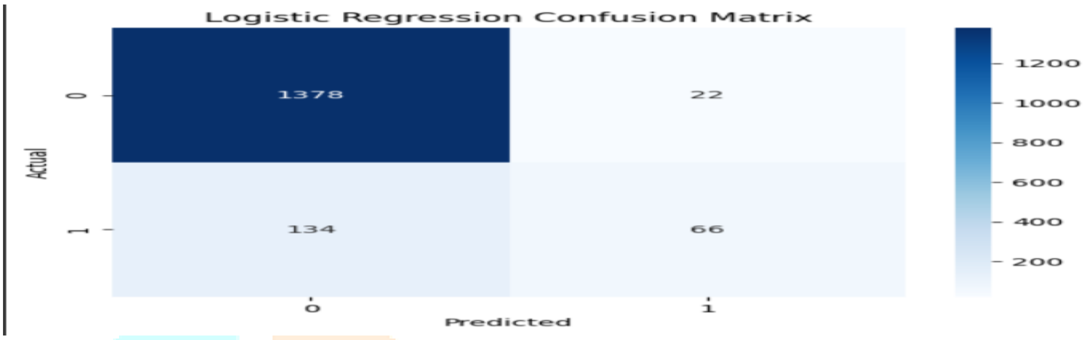


Fig 7- Logistic Regression

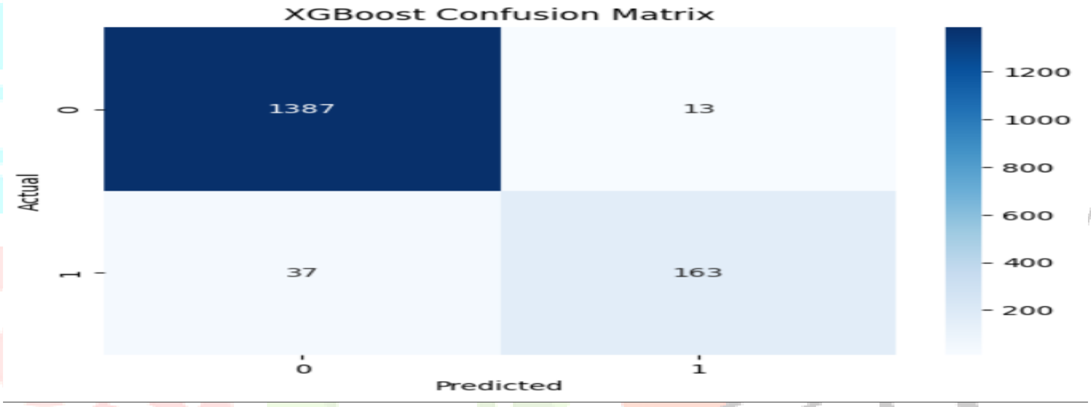


Fig 8- XG Boost Confusion Matrix

9.FUTURE SCOPE

With the growing concern over water pollution and increasing health risks due to contaminated water sources, the need for an intelligent and automated water quality testing system is more crucial than ever. The machine learning-based solution developed in this project addresses this problem directly. It can be extended into a real-time monitoring system capable of not only assessing current water samples but also verifying the reliability of water sources over time—much like how OTP systems are used regularly for secure access This approach can be used routinely to evaluate water quality from various sources, such as public water supplies, wells, or industrial outflows. Moreover, it can be applied to historical datasets to detect patterns or periods of unsafe water conditions, thereby helping in tracing contamination events and generating data-driven evidence for accountability or further investigation. Additionally, the proposed models can be optimized for better accuracy and speed, and new machine learning algorithms can be tested to improve performance. Such systems could eventually be integrated with IoT sensors for automated, continuous water quality surveillance in smart cities, rural areas, and industrial zones alike.

10. CONCLUSION

Although there are several water quality monitoring techniques available today, most are not capable of identifying unsafe water conditions in real time. Typically, these systems detect contamination after the water has already been consumed or distributed, which could lead to serious health risks. This delay occurs because unsafe water samples form only a small fraction of the total samples, making real-time detection a challenging task. Therefore, there is a need for a smart, cost-effective, and fast detection system that can identify unsafe water as soon as it is analyzed, enabling timely action and prevention of potential health hazards. The primary goal today is to build a highly accurate, sensitive, and real-time water quality detection system using machine learning, which can work efficiently across various environments—whether it's urban pipelines, rural wells, or industrial discharge systems. However, a major limitation of current techniques is their inconsistent performance across different datasets or environments. A model trained on one water source might not perform well on data from a different region due to variation in water composition and contamination types. In our balanced (undersampled) dataset, we observed that the model struggles to correctly classify a significant number of safe water samples, instead labeling them as unsafe. This can have major implications—for instance, if water that is actually safe is wrongly flagged as unsafe, it could lead to unnecessary panic or resource wastage, and damage public trust. To address this, further steps like outlier detection and removal will be applied to the dataset to see if the model's accuracy improves.In this project, machine learning techniques such as Logistic Regression, Support Vector Machine, Random Forest, and

XGBoost were used to predict water safety. The models were evaluated using metrics like accuracy, precision, recall (sensitivity), and F1-score. Based on the results, Random Forest and Logistic Regression emerged as the best-performing models for our water quality classification system, striking a balance between interpretability and predictive power.

ACKNOWLEDGEMENT

Kandhati Tulasi Krishna Kumar: Training & Placement Officer with 16 years' experience in training & placing the students into IT, ITES & Core profiles & trained more than 9,700 UG, PG candidates & trained more than 450 faculty through FDPs. Authored various books for the benefit of the diploma, pharmacy, engineering & pure science graduating students. He is a Certified Campus Recruitment Trainer from JNTUA, did his Master of Technology degree in CSE from VTA and in process of his Doctoral research. He is a professional in Pro-E, CNC certified by CITD He is recognized as an editorial member of IJIT (International Journal for Information Technology & member in IAAC, IEEE, MISTE, IAENG, ISOC, ISQEM, and SDIWC. He published 6 books, 55 articles in various international journals on Databases, Software Engineering, Human Resource Management and Campus Recruitment & Training.

Gadula Noshna is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Artificial intelligence K.Bhargavi has taken up his PG project on Water quality prediction using machine learning and published the paper in connection to the project under the guidance of Kandhati Tulasi Krishna Kumar Nainar, Assistant Professor, Training & Placement Officer, SVPEC

REFERENCES

1. Mosavi, F. S. Hosseini, B. Choubin, M. Goodarzi and A. A. Dineva, "Groundwater Salinity Susceptibility Mapping Using Classifier Ensemble and Bayesian Machine Learning Models," *in IEEE Access*, vol. 8, pp. 145564–145576, 2020, doi: 10.1109/ACCESS.2020.3014908"Performance evaluation of deep learning and boosted trees for BITCOIN closing price prediction," by Azeez A. Oyedele et al. 2023: 119233 Expert Systems with Applications 213

2. O. Al-Sulttani, M. Al-Mukhtar, A. B. Roomi, A. A. Farooque, K. M. Khedher and Z. M. Yaseen,"Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction," *in IEEE Access*, vol. 9, pp. 108527–108541, 2021, doi: 10.1109/ACCESS.2021.3100490.Making BITCOIN Price Predictions Using Machine Learning Ireland, Dublin, IEEE 2018, Sean McNally, Jason Roche, and Simon Caton.

3. Hongfang Lu, Xin Ma,"Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, Volume 249, 2020, 126169, ISSN 0045-6535, https://doi.org/10.1016/j.chemosphere.2020.126169Bi-LSTM Network used to predict the price of BITCOIN. International Conference on Informatics and Computer Communication in 2021. P. Nithyakani and others (ICCCI). IEEE, 2021.

4. "Hadi Mohammed, Hoese Michel Tornyeviadzi, Razak Seidu, "Emulating process-based water quality modelling in water source reservoirs using machine learning," *Journal of Hydrology*, Volume 609, 2022, https://doi.org/10.1016/jjhydroL2022.127675Jacob Coburn and Sara C. Pryor's "Projecting Future Energy Production from Operating Wind Farms in North America: Part II: Statistical Downscaling." Journal of Applied Meteorology and Climatology 62.1; 2023: 81–101.

5. Botao Chen, Xi Mu, Peng Chen, Biao Wang, Jaewan Choi, Honglyun Park, Sheng Xu, Yanlan Wu, Hui Yang, "Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data," *Ecological Indicators*, Volume 133, 2021, https://doi.org/10.1016/j.ecolind.2021.108434"Spatial prediction of groundwater potential and driving factor analysis based on deep learning and geographical detector in a dry endorheic basin," by Wang, Zitao, Jianping Wang, and Jinjun Han 109256 is the ecological indicator number for 2022.

6. Hye Won Lee, Min Kim, Hee Won Son, Baehyun Min, Jung Hyun Choi, "Machine-learning-based water quality management of river with serial impoundments in the Republic of Korea," *Journal of Hydrology: Regional Studies*, Volume 41, 2022, https://doi.org/10.1016/j.ejrh.2022.101069Chandra, MunipalliSasi, R. Sumathi, and J. Jeyaranjani. "Analysis of Predicting BITCOIN Price using Deep Learning Technique." *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, 2022.

7. Mashud Rana, Ashfaqur Rahman, Joel Dabrowski, Stuart Arnold, John McCulloch, Bruno Pais, “Machine learning approach to investigate the influence of water quality on aquatic livestock in freshwater ponds,” *Biosystems Engineering*, Volume 208, 2021, https://doi.org/10.1016/j.biosystemseng.2021.05.017

8. Ali El Bilali, Abdeslam Taleb, “Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment,” *Journal of the Saudi Society of Agricultural Sciences*, Volume 19, Issue 7, 2020, https://doi.oig/10.1016Zj.jssas.2020.08.001

9. Luka Grbcic, Sinisa Druzeta, Goran Mausa, Tomislav Lipic, Darija Vukic Lusic, Marta Alvir, Ivana Lucin, Ante Sikirica, Davor Davidovic, Vanja Travas, Daniela Kalafatovic, Kristina Pikelj, Hana Fajkovic, Toni Holjevic, Lado Kranjcevic, “Coastal water quality prediction based on machine learning with featureinterpretationand spatio-temporal analysis,” *Environmental Modelling A Software*, Volume 155, 2022, https://doi.org/10.1016/j.envsoft.2022.105458

10. Kangyang Chen, Hexia Chen Chuanlong Zhou, Yichao Huang, Xiangyang Qi, Ruqin Shen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng Hongqiang Ren, “Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data,” *Water Research*, Volume 171, 2020, https://doi.org/10.1016/j.watres.2019.115454

11. Islam Khan, d.S., Islam, N., Uddin, J., Islam, S., Nasir, M.K., “Water Quality Prediction and Classification Based on Principal Component Regression and Gradient Boosting Classifier Approach,” *Journal of King Saud University - Computer and Information Sciences (2021)*, doi: https://doi.org/10.1016/j.jksuci.2021.06.003

12. D. Venkata Vara Prasad, P. Senthil Kumar, “*Automating water quality analysis using ML and auto ML techniques*,” https://doi.org/10.1016/j.envres.2021.111720

13. Juna, Afaq, Muhammad Umer, Saima Sadiq, Hanen Karamti, Ala’ Abdulmajid Eshmawi, Abdullah Mohamed, and Imran Ashraf “Water Quality Prediction Using KNN Imputer and Multilayer Perceptron,” *Water* 14, no. 17 : 2592, https://doi.org/10.3390/wl4172592

14. Hao Liao, Wen Sun,“Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method,” *Procedia Environmental Sciences*, Volume 2, 2010, Pages 970–979, https: //doi. org/ 10.1016/j.proenv.2010.10.109

15. Ahmed, Umair & Mumtaz, Rafia & Anwar, Hirra & Shah, Asad & Irfan, Rabia & Garcia-Nieto Jose. (2019).“Efficient Water Quality Prediction Using Supervised Machine Learning,” *Water*. 11. 2210. 10.3390/w11112210.