International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Intelligent Machinelarning System for Water Quality and Potability Prediction

K. Devi Venkata Sai Harshini[1], Sk. Rabbani[2], R. Sri Mahalakshmi[3], C. Lakshmi Rajeswari[4], Dr. Koppula Chinabusi[5]

[1,2,3,4]Final Year Students, Tirumala Institue of Technology and Sciences, Narasaraopet
[5]Professor, HOD, Department of Computer Science,

Abstract
Access to safe drinking water is essential for human health and environmental sustainability. Traditional methods of water quality assessment rely on laboratory-based testing, which is time-consuming, costly, and unsuitable for real-time monitoring. This paper presents a machine learning-based approach for predicting water potability using physicochemical parameters such as pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. Multiple supervised learning algorithms, including Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Classifier (SVC), are trained and evaluated to determine the most effective model. Based on performance comparison, SVC is selected due to its superior accuracy and generalization capability. The trained model is integrated into a web-based application that enables users to input water quality parameters and obtain real-time predictions along with confidence scores. The proposed system provides a fast, cost-effective, and reliable solution for water quality assessment and supports improved decision-making in environmental monitoring.

Index Terms: Machine Learning, K-nearest neighbor, Random Forest, Support Vector Classifier (SVC), Potability Analysis, Data Preprocessing, Feature Scaling, Classification Algorithms, Environmental Monitoring, Web-Based Application, Real-Time Prediction, Confidence Score

1. INTRODUCTION

Access to clean and safe drinking water is a fundamental requirement for human health and environmental sustainability. However, rapid industrialization, urbanization, and population growth have significantly affected water quality, leading to increased contamination of water resources. Consumption of polluted water can cause serious health issues, including waterborne diseases, making regular monitoring of water quality essential. Traditional water quality assessment methods are primarily based on laboratory testing of water samples. Although these methods provide accurate results, they are time-consuming, costly, and require specialized equipment and skilled personnel. Moreover, such approaches are not suitable for real-time monitoring or large-scale implementation, which limits their effectiveness in ensuring continuous water quality assessment. With the advancement of data science and artificial intelligence, machine learning techniques have emerged as a promising solution for water quality prediction. Machine learning models can analyze historical water quality data and identify patterns that help in classifying water as potable or non-potable. These models offer faster processing, reduced cost, and the ability to handle

complex relationships among multiple parameters. In this work, a machine learning-based approach is proposed for predicting water potability using key physicochemical parameters such as pH, hardness, dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. Multiple classification algorithms are evaluated, and the Support Vector Classifier (SVC) is selected as the most effective model based on performance metrics. The selected model is further deployed as a web-based application that enables users to input parameters and obtain real-time predictions along with confidence scores.

The remainder of this paper is organized as follows: Section II presents the literature survey, Section III describes the system analysis, Section IV discusses the system design, and subsequent sections present implementation, results, and conclusions.

2. Literature survey

Water quality prediction has gained significant attention due to the increasing demand for safe drinking water and the limitations of traditional laboratory testing methods. Early approaches mainly relied on statistical analysis and manual evaluation, which were time-consuming and less efficient

1. Water quality prediction has gained significant attention due to the increasing demand for safe and sustainable water resources. Traditional laboratory-based methods, although accurate, are time-consuming and costly, making them unsuitable for real-time monitoring. To address these limitations, researchers have explored machine learning techniques for efficient and automated water quality assessment.
2. Early studies focused on statistical and rule-based approaches for water quality analysis. However, these methods were limited in handling complex and nonlinear relationships among water quality parameters. With the advancement of machine learning, models such as Logistic Regression and Decision Trees were introduced, providing improved predictive capabilities. Nevertheless, these models often faced challenges when dealing with high-dimensional and nonlinear datasets .
3. To enhance prediction accuracy, ensemble learning techniques such as Random Forest have been widely adopted. Random Forest improves robustness and reduces overfitting by combining multiple decision trees.Additionally, Support Vector Machines (SVM) have shown strong performance in classification tasks due to their ability to model nonlinear decision boundaries and handle high-dimensional data effectively. Several studies have reported that SVM outperforms traditional classifiers in water quality prediction.
4. Other machine learning approaches, including K-Nearest Neighbors (KNN) and Naïve Bayes, have also been applied in this domain. KNN is simple and effective but sensitive to noise and feature scaling, whereas Naïve Bayes is computationally efficient but assumes feature independence.
5. Despite these advancements, most existing works focus primarily on improving model accuracy and lack real-time deployment and user-friendly interfaces. The proposed system addresses these limitations by comparing multiple machine learning models, selecting the Support Vector Classifier (SVC) based on performance, and deploying it as a web-based application that provides real-time predictions along with confidence scores.

3. Existing System

The existing system for water quality assessment mainly relies on laboratory-based testing methods. Water samples are collected and analyzed using specialized equipment to measure various parameters. Although

these methods provide accurate results, they require skilled personnel and significant time for analysis. However, the traditional approach has several limitations, including high cost, time consumption, and lack of real-time monitoring. The process is not suitable for large-scale or continuous water quality assessment. These drawbacks highlight the need for an automated and efficient system for water quality prediction.

4. Proposed System

The proposed system presents a machine learning-based approach for predicting water potability using physicochemical parameters. The system is designed to overcome the limitations of traditional water quality assessment methods by providing a fast, automated, and reliable solution. It utilizes historical water quality data containing attributes such as pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity to train predictive models.

The system begins with a data preprocessing phase, where missing values are handled and feature scaling techniques are applied to normalize the data. This step ensures that all input parameters are consistent and suitable for machine learning algorithms. Multiple classification models, including Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Classifier (SVC), are trained and evaluated using standard performance metrics such as accuracy.

Based on comparative analysis, the Support Vector Classifier (SVC) is selected as the final model due to its superior accuracy and ability to handle nonlinear relationships among features. The trained model is then integrated into a web-based application, enabling real-time interaction with users. The system accepts input parameters from users, processes them using the same preprocessing techniques, and generates predictions indicating whether the water is potable or non-potable.

In addition to classification, the system provides a confidence score for each prediction, enhancing transparency and user trust. The web interface is designed to be simple and user-friendly, allowing users with minimal technical knowledge to operate the system.Overall, the proposed system offers an efficient, scalable, and cost-effective solution for water quality prediction and can be extended in the future with real-time data integration and advanced machine learning techniques.

5. DATA

The dataset used in this study is the Water Potability Dataset, which is widely utilized for analyzing and predicting water quality using machine learning techniques. It contains a comprehensive set of physicochemical parameters that influence the potability of water. Each record in the dataset represents a water sample with measured attributes and a corresponding target label indicating whether the water is safe for human consumption.

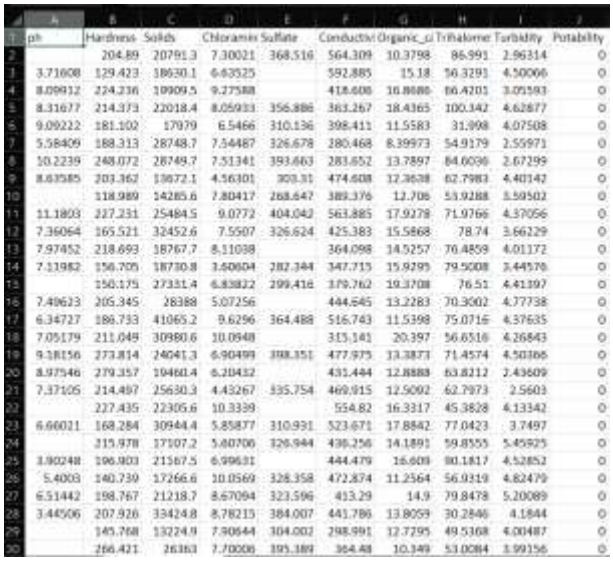
The dataset includes several important features such as pH, which indicates the acidity or alkalinity of water; hardness, which represents the concentration of calcium and magnesium salts; total dissolved solids (TDS), which measure the combined content of inorganic and organic substances; chloramines, used as disinfectants; sulfate, which can affect taste and health; conductivity, indicating the ability of water to conduct electricity; organic carbon, reflecting organic pollution; trihalomethanes, which are by-products of water disinfection; and turbidity, which measures the clarity of water. These parameters collectively provide a detailed representation of water quality conditions.

The target variable, Potability, is a binary classification label, where a value of 1 indicates that the water is potable (safe to drink), and 0 indicates that it is non-potable. The dataset may contain missing values in certain attributes due to incomplete measurements or data collection issues. To address this, data

preprocessing techniques such as mean imputation are applied to fill missing values, ensuring that the dataset remains consistent and usable for model training.

Before training the machine learning models, feature scaling techniques such as standardization are applied to normalize the range of input variables. This step is crucial because it ensures that all features contribute equally to the learning process, especially for algorithms like Support Vector Classifier (SVC) that are sensitive to feature scales. The dataset is then divided into training and testing subsets, typically using an 80:20 ratio, to evaluate model performance and generalization ability.

Overall, the dataset plays a critical role in the success of the proposed system by providing relevant and structured information required for accurate prediction. Its diverse set of features enables the machine learning models to capture complex relationships between water quality parameters and potability, leading to reliable and efficient prediction outcomes.



ph	Hardness	Solids	Chloramin	Sulfate	Conductivi	Organic_ci	Trihalome	Turbidity	Potability
	204.89	20791.3	7.30021	368.516	564.309	10.3798	86.991	2.96314	0
3.71608	129.423	18630.1	6.63525		592.885	15.18	56.3291	4.50066	0
8.09912	224.236	19909.5	9.27588		418.606	16.8686	66.4201	3.05593	0
8.31677	214.373	22018.4	8.05933	356.886	363.267	18.4365	100.342	4.62877	0
9.09222	181.102	17979	6.5466	310.136	398.411	11.5583	31.998	4.07508	0
5.58409	188.313	28748.7	7.54487	326.678	280.468	8.39973	54.9179	2.55971	0
10.2239	248.072	28749.7	7.51341	393.663	283.652	13.7897	84.6036	2.67299	0
8.63585	203.362	13672.1	4.56301	303.31	474.608	12.3638	62.7983	4.40142	0
	118.989	14285.6	7.80417	268.647	389.376	12.706	53.9288	3.59502	0
11.1803	227.231	25484.5	9.0772	404.042	563.885	17.9278	71.9766	4.37056	0
7.36064	165.521	32452.6	7.5507	326.624	425.383	15.5868	78.74	3.66229	0
7.97452	218.693	18767.7	8.11038		364.098	14.5257	70.4859	4.01172	0
7.13982	156.705	18730.8	3.60604	282.344	347.715	15.9295	79.5008	3.44576	0
	150.175	27331.4	6.83822	299.416	379.762	19.3708	76.51	4.41397	0
7.49623	205.345	28388	5.07256		444.645	13.2283	70.3902	4.77738	0
6.34727	186.733	41065.2	9.6296	364.488	516.743	11.5398	75.0716	4.37635	0
7.05179	211.049	30980.6	10.0948		315.141	20.397	56.6516	4.26843	0
9.18156	273.814	24041.3	6.90499	388.351	477.975	13.3873	71.4574	4.50386	0
8.97546	279.357	19460.4	6.20432		431.444	12.8888	63.8212	2.43609	0
7.37105	214.497	25630.3	4.43267	335.754	469.915	12.5092	62.7973	2.5603	0
	227.435	22305.6	10.3339		554.82	16.3317	45.3828	4.13342	0
6.66021	168.284	30944.4	5.85877	310.931	523.671	17.8842	77.0423	3.7497	0
	215.978	17107.2	5.80706	326.944	436.256	14.1891	59.8555	5.45925	0
3.90248	196.903	21567.5	6.99631		444.479	16.609	80.1817	4.52852	0
5.4003	140.739	17266.6	10.0569	328.358	472.874	11.2564	56.9319	4.82479	0
6.51442	198.767	21218.7	8.67094	323.596	413.29	14.9	79.8478	5.20089	0
3.44506	207.926	33424.8	8.78215	384.007	441.786	13.8059	30.2846	4.1844	0
	145.768	13224.9	7.90644	304.002	298.991	12.7295	49.5368	4.00487	0
	266.421	26163	7.70008	395.389	364.48	10.349	53.0084	3.99156	0

Fig. 1. The Ten Feature for Assessing the Potability of Water.

A. Data Exploration and Preparation

As part of data pre-processing, the data were converted into float after being in a string. In addition, the data that were in repetition were deleted, so only the necessary data were kept, see Fig. 2.

Initially, we will check whether there are NULL values or not. This is important to ensure that the algorithm can run smoothly without any missing data since null values indicate

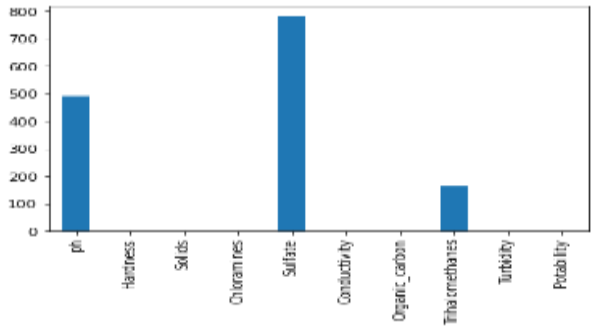


Fig. 2. The Value of Necessary Data Needed for our Model.

missing data. Furthermore, the algorithm can obtain more accurate results when the null values are replaced. As can be seen in the image below, pH, Sulfate, and Trihalomethanes have NULL values. As a

solution, the null values are usually replaced by the average or mean of the specific category. After that the information in the dataset is checked again.

The mean value is calculated by measuring the sum of the available values divided by the total number of values in the categories. This mean calculation is used to handle the missing data relative to NULL values, see Fig. 3.

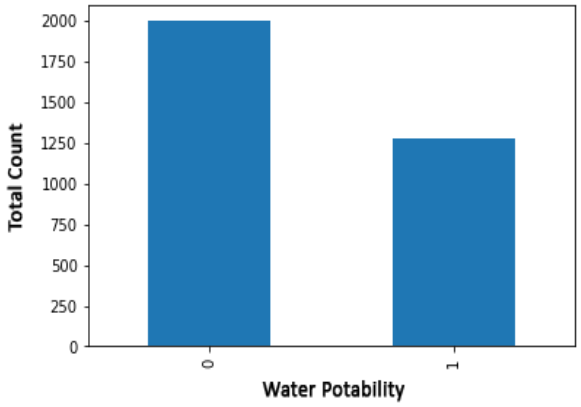


Fig. 3. The Sum of the Available Values for Water Potability in the Dataset: 0 Water is not Potable, 1 Water is Potable

6. RESULT

The performance of the proposed water quality prediction system is evaluated using multiple supervised machine learning algorithms, including Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC). The dataset is divided into training and testing subsets, and all models are trained on the preprocessed data to ensure fair comparison. The evaluation is carried out using standard performance metrics such as accuracy and classification effectiveness.

Experimental results indicate that the Support Vector Classifier (SVC) outperforms the other models in terms of prediction accuracy and generalization capability. This superior performance can be attributed to the ability of SVC to handle nonlinear relationships between water quality parameters using kernel functions. While Random Forest and Decision Tree models also demonstrate good performance due to their ensemble and hierarchical structure, their accuracy is slightly lower compared to SVC. Logistic Regression and Naïve Bayes show comparatively moderate performance, whereas KNN is affected by sensitivity to feature scaling and noise in the dataset.

The results further highlight the importance of data preprocessing techniques in improving model performance. Handling missing values and applying feature scaling significantly enhance the accuracy of all models, particularly SVC, which is sensitive to feature magnitudes. The trained model is capable of accurately classifying water samples as potable or non-potable based on the given input parameters.

In addition to classification, the system provides a confidence score for each prediction using probability estimates. This feature enhances the interpretability and reliability of the system, allowing users to understand the certainty of the prediction. The model is integrated into a web-based application that enables real-time prediction, making the system practical for real-world usage.

Overall, the experimental results demonstrate that the proposed system is effective, reliable, and suitable for water quality prediction. The combination of accurate machine learning models, proper preprocessing, and real-time deployment ensures that the system can be used as a valuable tool for environmental

monitoring and decision-making.

Fig. 4. Giving inputs in web page.

After providing the input click “predict” and then user will get the output as follows

Fig. 5. Output Display

Based on the given input parameters output will be displayed

Machine Learining Algorithm Accuracy:

The performance of various machine learning algorithms is evaluated based on their accuracy in predicting water potability. Multiple classification models, including Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC), are trained and tested on the preprocessed dataset.

Among all the models, the Support Vector Classifier (SVC) achieves the highest accuracy, indicating its effectiveness in handling nonlinear relationships between water quality parameters. Random Forest also demonstrates strong performance due to its ensemble learning approach, while Decision Tree provides moderate accuracy. Logistic Regression and Naïve Bayes show comparatively lower performance due to their assumptions and limitations. KNN performance is affected by sensitivity to feature scaling and noise. Overall, the comparison of model accuracies highlights that SVC is the most suitable algorithm for this problem, providing better generalization and reliable prediction results.

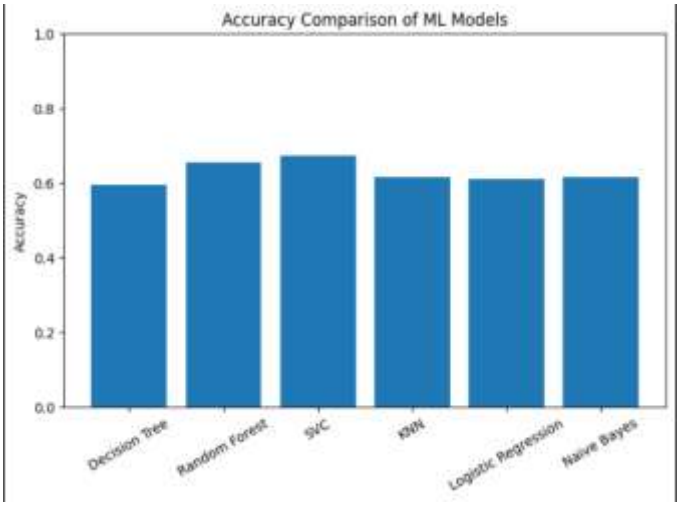
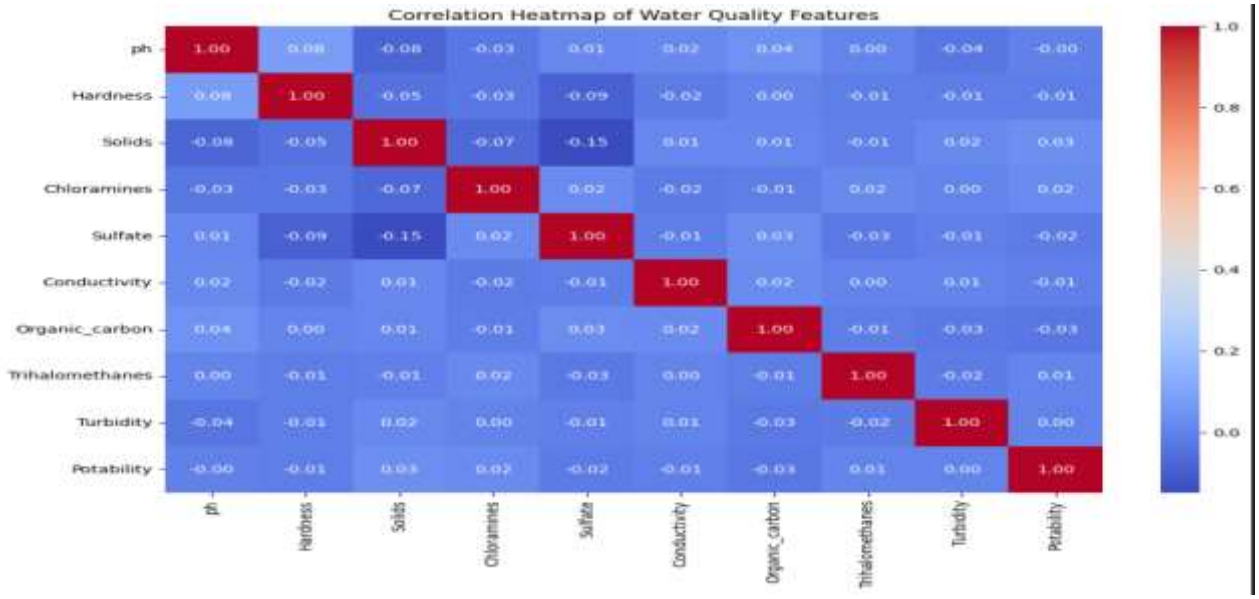


Fig. 6. Bargraph for accuracy of machine learning algorithms.

Accuracy In Terms Of Numbers:
Decision Tree= 0.594512
Random Forest= 0.653963
SVC= 0.670731
KNN= 0.614329
Logistic Regression= 0.609754
Naïve Bayes= 0.614329

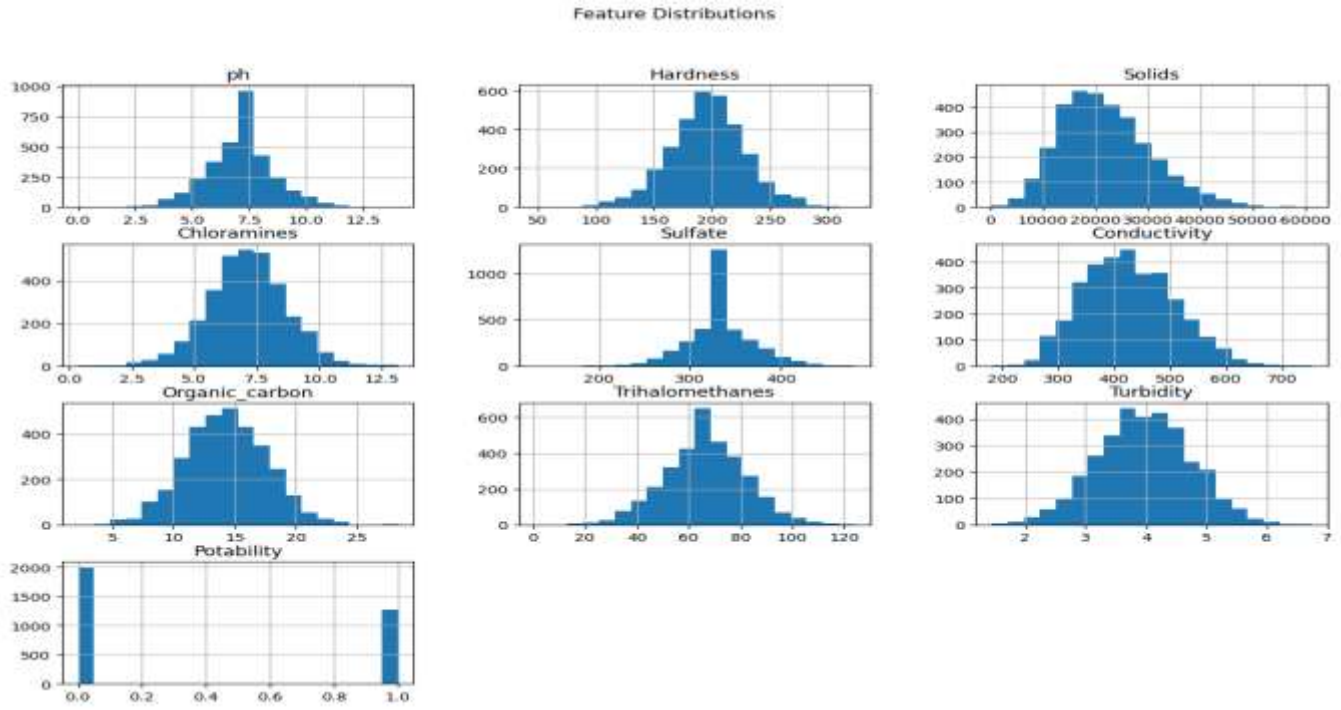
In the context of water quality prediction, True Positive represents the number of correctly predicted potable water samples, while True Negative indicates correctly predicted non-potable samples. False Positive refers to cases where non-potable water is incorrectly classified as potable, and False Negative represents potable water incorrectly classified as non-potable.

The confusion matrix helps in understanding not only the accuracy of the model but also its ability to correctly identify both classes. It is particularly useful for identifying misclassification errors and improving model performance. The results show that the selected Support Vector Classifier (SVC) model achieves a high number of correct predictions, indicating it effectiveness and reliability in water potability classification



In the context of water quality prediction, True Positiverepresents the number of correctly predicted potable water samples, while True Negative indicates correctly predicted non-potable samples. False Positive refers to cases where non-potable water is incorrectly classified as potable, and False Negative represents potable water incorrectly classified as non-potable.

The confusion matrix helps in understanding not only the accuracy of the model but also its ability to correctly identify both classes. It is particularly useful for identifying misclassification errors and improving model performance. The results show that the selected Support Vector Classifier (SVC) model achieves a high number of correct predictions, indicating its effectiveness and reliability in water potability classification.



7. CONCLUSION

In this paper, a machine learning-based system for water quality prediction has been successfully developed and evaluated. The study focuses on analyzing various physicochemical parameters to determine the potability of water. Traditional laboratory-based testing methods, although accurate, are time-consuming and expensive, which limits their practical use for real-time monitoring. The proposed system addresses these limitations by providing an automated and efficient prediction mechanism using machine learning techniques.

A comprehensive analysis of multiple classification algorithms, including Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Classifier (SVC), has been performed. Among these models, SVC demonstrated superior performance in terms of accuracy and generalization capability, making it the most suitable choice for this application. The effectiveness of the model is further enhanced by applying appropriate data preprocessing techniques such as handling missing values and feature scaling, which significantly improve prediction accuracy.

The developed system is implemented as a web-based application, enabling users to input water quality parameters and obtain real-time predictions along with confidence scores. This enhances the usability and accessibility of the system, making it suitable for practical deployment in environmental monitoring and

decision-making processes. The results indicate that the system is capable of providing reliable and consistent predictions, thereby supporting efforts to ensure safe drinking water.

Despite its effectiveness, the system has certain limitations, such as dependency on the quality and size of the dataset. Future work can focus on incorporating real-time data acquisition through IoT-based sensors, expanding the dataset, and exploring advanced machine learning and deep learning techniques to further improve performance. Overall, the proposed system represents a significant step toward efficient and intelligent water quality assessment.

REFERENCES

1. S. K. Yadav and S. Singh, "Water quality prediction using machine learning techniques," International Journal of Engineering Research & Technology, vol. 9, no. 5, pp. 123–128, 2020.
2. A. K. Sharma and P. Sharma, "Analysis of water quality using data mining techniques," International Journal of Computer Applications, vol. 150, no. 12, pp. 10–15, 2019.
3. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
4. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
5. T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
6. N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, vol. 29, no. 2–3, pp. 131–163, 1997.
7. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
8. M. Zhu, J. Liu, and H. Wang, "Machine learning techniques for water quality prediction: A review," Eco-Environment & Health, vol. 1, no. 1, pp. 1–12, 2022.
9. X. Yan, "Application of machine learning in water quality assessment: A review," Journal of Marine Science and Engineering, vol. 12, no. 1, pp. 1–20, 2024.
10. M. Bagheri, A. Mohammadi, and M. Zare, "Artificial intelligence methods for water quality prediction," Journal of Hydrology, vol. 585, pp. 124–135, 2020.
11. H. Liao, Y. Sun, and Z. Wang, "Water quality prediction using decision tree method," Environmental Monitoring and Assessment, vol. 165, no. 1–4, pp. 89–98, 2010.
12. J. Chou, Y. Chen, and C. Chen, "Predicting water quality using machine learning techniques," Environmental Modelling & Software, vol. 101, pp. 1–12, 2018.
13. U. J. Nwankwo and P. O. Okafor, "Artificial neural network model for water quality prediction," Applied Water Science, vol. 11, no. 3, pp. 1–10, 2021.
14. H. Mohammed, R. Ibrahim, and S. Hassan, "Predictive modeling of water quality using machine learning," Water Resources Management, vol. 32, no. 5, pp. 1501–1515, 2018.
15. T. Li, X. Zhang, and Y. Wang, "Aquaculture water quality prediction using machine learning," Ecological Informatics, vol. 68, pp. 101–110, 2022.