iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.59747

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Potable Water Quality Prediction: By Artificial Intelligence Techniques with Advanced Machine Learning Algorithm's

Vijendra S N[1], Prashant[2], Jayaprakash M[3], Ananya R [4], Shivashankar N[5]
[1]*Assistant Professor, Dept of CSE Impact College of Engineering and Applied sciences, Bangalore, Affiliated to VTU*
[2,3,4,5] *Students, Dept of CSE Impact College of Engineering and Applied sciences, Bangalore, Affiliated to VTU*

Abstract: ***Water is necessary for humans to survive, and everyone's health depends on maintaining the quality of the resource. Drinking polluted water can put one's health at risk, raising the chances of contracting diseases like cholera and other waterborne infections. By predicting the water's quality, 'machine learning algorithms' have developed into beneficial tools for quickly and reliably monitoring water supplies. Many forecasting techniques are the main subject of this study. This project aims to estimate water potability using various algorithms by forecasting the physicochemical characteristics of water samples taken from the Drinking Water dataset on Kaggle. To find the potability of drinking water, we use a variety of methods, including 'random forest', 'logistic regression', 'decision tree', 'SVM', 'AdaBoost', and 'KNN'. There is hence a strong chance that the investigation will yield precise data regarding the quality of the water.***

Keywords: Water Potability, 'Machine Learning Algorithms', Physicochemical Characteristics, Forecasting techniques, Drinking Water Dataset.

I. INTRODUCTION

When it comes to India, It's been demonstrated by a disturbing estimate that around 37.7 million people, including children, get waterborne illnesses every year. Unbelievably, industrial and residential pollutants have tainted around seventy percent water supply, depriving between 20% and 80% comprises both rural and urban residents, respectively, of clean drinking water. Furthermore, 49 Million individuals throughout the world suffer due to the serious issues the globe is facing about water shortages, and 28 declining water quality. Alarming data from the World Health Organization for the year 2018 report says that people consuming fecal matter contaminated water is about 2 billion. Therefore, guaranteeing universal access is necessary to ensure sustainable development, encourage a healthy lifestyle and end poverty.

Predicting water potability plays a critical role in places with poor water treatment infrastructure, including developing countries and rural areas. Conventional approaches to water quality monitoring, which entail pricey and time-consuming statistical and laboratory studies, are ineffective. As such, a more economical and efficient substitute is desperately needed. This investigation aims to present and evaluate a machine learning-based method for real-time, accurate water potability prediction. The use of 'machine learning algorithms' to forecast water quality has advanced significantly over the past several years, improving the efficacy and accuracy of monitoring. To forecast the potability of water, an array of classification methods may be used, such as 'decision trees', 'SVM', 'RF', and 'KNN' approaches.

This article primarily focuses on using physicochemical features to determine drinking water's potability. The intention behind the project is to produce a model that can deliver precise and timely data on the quality of drinking water, empowering managers of water resources and policymakers to take preventative action and guarantee that the general population has access to safe drinking water. The study also attempts to evaluate the efficacy and precision of several water quality prediction systems.

The following is a list of several concerns that this research aims to resolve:

1) The WHO's guidelines for defining potable freshwater parameters might be interpreted incorrectly.
2) A exists lack of applications for water potability and water quality prediction.
3) The clinical method currently used to predict drinkable water is time-consuming and inefficient.
4) It's possible that rural residents are unaware of important awareness elements that could have an impact on their access to safe drinking water.

II. LITERATURE SURVEY

Predicting water quality and evaluating portability are important fields of study that contribute to providing clean potable water to communities throughout the globe. The overview explores many works investigating machine learning techniques in this field. Gradient Boost and Random Forest emerged as the top-performing explainable AI techniques in Patel et al.'s methodology, which addressed class imbalance. When Uddin et al. assessed a water quality indicator model, With an accuracy rate of more than 67%, Support Vector Machines (SVM) fared better than the other techniques. In Uddin et al.'s evaluation of a water quality indicator model, support vector machines (SVM) showed better accuracy. Furthermore, Tufail et al.'s investigation on water potability forecasting in Pakistan revealed that SVM outperformed random forests and decision trees Regarding effectiveness. When taken as a whole, these findings highlight how crucial machine learning is to improving water quality prediction.

III.METHODOLOGY

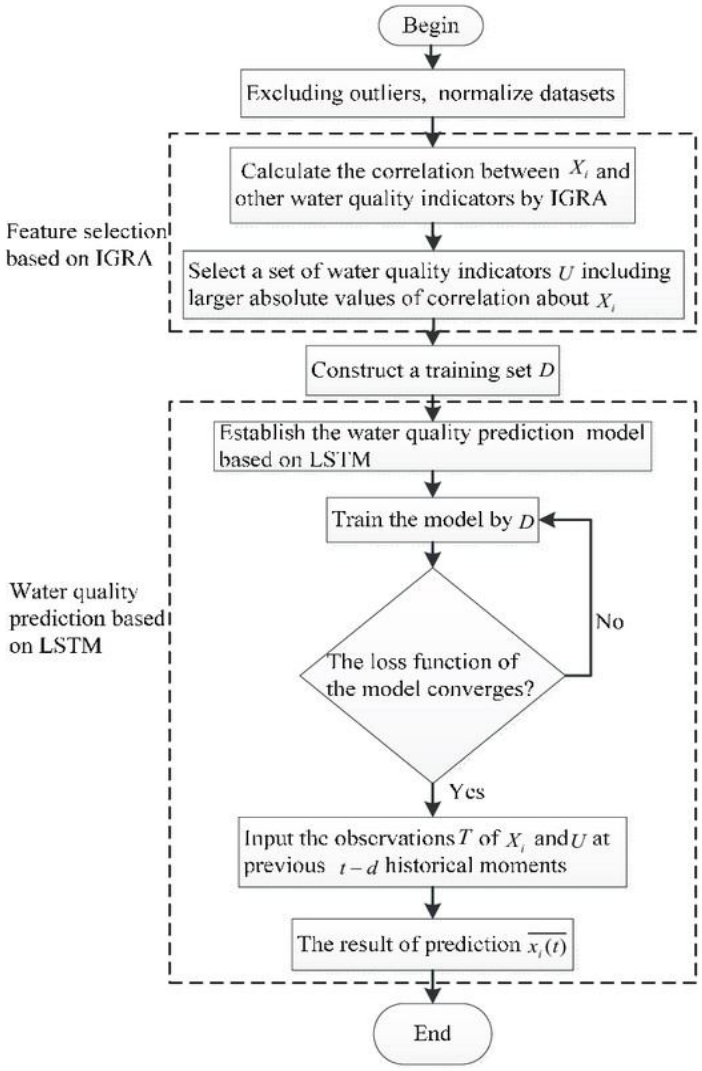


Figure 1: Workflow for water potability prediction model

Ensuring the security of potable water involves a complex process that considers various 'physical', 'chemical', and biological factors. New research indicates that 'ML' algorithms can accurately predict water quality and determine its potability. This investigation attempts to establish a predictive model with greater accuracy in water potability using ML models. Models can help in effective water management and ensure a steady supply of safe drinking water in communities. Figure 1 demonstrates the research progress, which includes data collection, preparation, handling of outliers and missing values, data normalization, model creation, performance evaluation, and fine-tuning of hyperparameters.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

A. Collecting Data

The primary source of evidence for this investigation was a Kaggle dataset that was made accessible to the public. This dataset contains 2293 water quality findings that came from a variety of locations. It also includes a target feature called portability, which is employed in making predictions using several types of ML algorithms, along with nine different physicochemical parameters: pH, hardness, solids, chloramines, sulfates, trihalomethanes, organic carbon, conductivity, and turbidity. Every water quality parameter recommended by the EPA and WHO has standard values shown in Table 1.

Table. 1: Drinkable Water Quality Standards

Parameters	Unit	Standards
pH	Range 0-14	6-8
Organic Carbon	mg/L	2
Chloramines	Ppm	4
Turbidity	NTU	5
Trihalomethanes	µg/L	80
Sulfate	mg/L	250
Hardness	mg/L	300
Conductivity	µS/cm	500
Solids	mg/L	1000

1) Data Distribution

To put the potable water requirement in context, the research analyzes 10 distinct water quality factors and presents specific statistics. The associations between various characteristics in a dataset are shown in a correlation heatmap. Understanding the direction and One benefit is that it can reveal the level of correlation between variables. The 'Correlation Coefficient' varies from -1 to 1, indicates the level of the linear relationship between two variables. The variables are said to travel in the same direction if the 'correlation coefficient' is between 0 and 1. The variables, on the other hand, move in opposition to one another when the correlation value is negative, it goes from -1 to 0. The value of 0 indicates that there is no connection between the variables. Within this research, the link between 10 water quality measures was examined using a correlation heatmap. The heatmap demonstrates the existence of a negative correlation of -0.14 between sulfate and solids and a positive correlation of 0.086 between pH and solids. An in-depth examination is displayed in Figure 2.

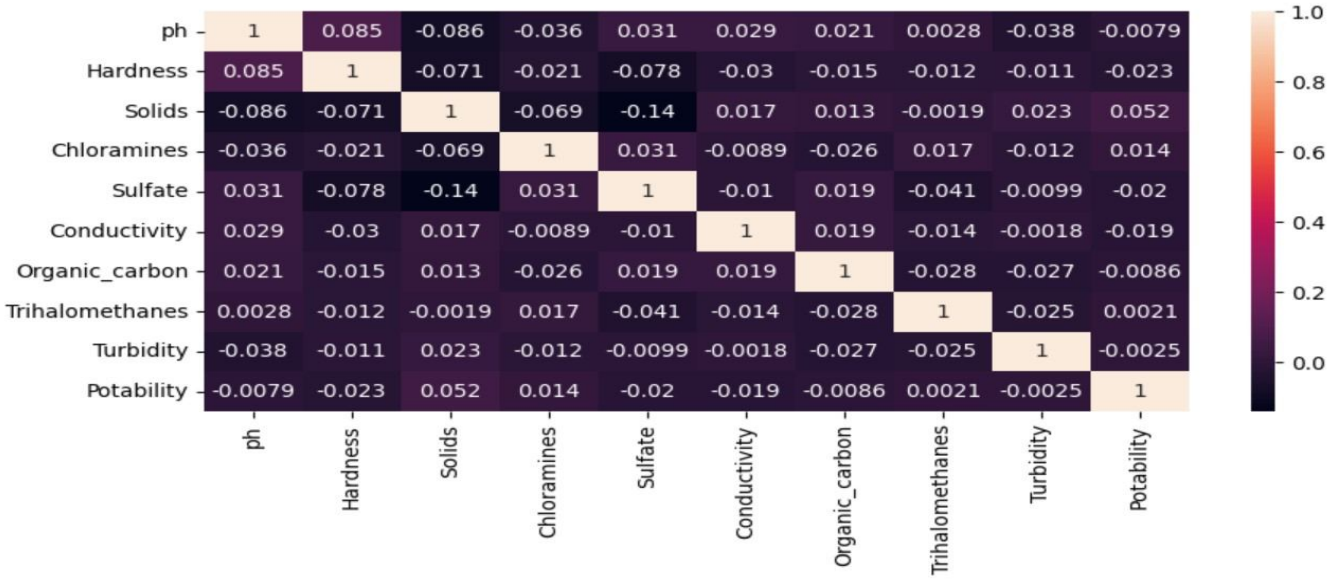


Figure 2: Correlation Heatmap

Figure 3 indicates that a sizable percentage of the specimens that were gathered—roughly 60%—fit into the category of not potable, indicating that they aren't fit for human consumption. Poor water quality may be caused by several things, such as manmade sources like sewage disposal, industrial effluents, and agricultural drainage, moreover natural events like erosion and climate change.

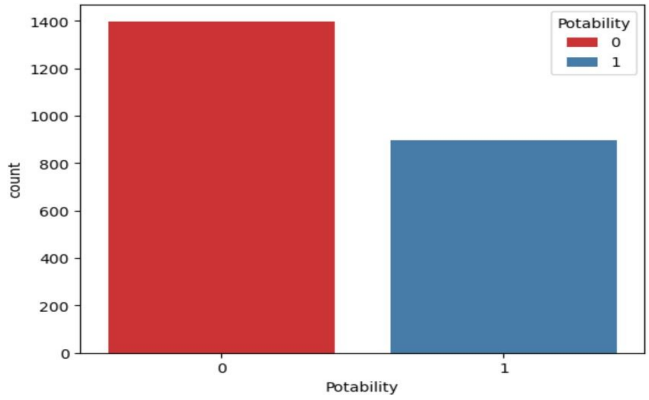


Figure 3: Potability feature distribution

B. Preprocessing

Data preparation is an important process in machine learning that ensures raw data is appropriately formatted for analysis. It involves selecting relevant features, handling outliers, and filling in missing information, and the data is scaled to fit a uniform range. Feature selection is also crucial, as it helps enhance ' model performance' by avoiding overfitting and decreasing complexity. Scaling features are also necessary, as they enhance the ‘efficiency’ of specific algorithms. Lastly, splitting the dataset into two sets: training and testing, guarantees proper model testing and learning.

C. Model Selection

The model selection procedure plays an important part in the analysis. It involves choosing the best machine-learning technique for the dataset and issue. In this study, the following algorithms are used:

1) ‘Logistic Regression’

'Logistic regression' is a popular algorithm used for predicting the drinkability of water, as it is effective in binary classification problems. It determines the probability that a water sample is safe to drink based on specific input features. Nevertheless, 'logistic regression' might not be as efficient in handling high-dimensional data once there's an additional predictor variable than observations. Additionally, it may overlook interaction effects since it presupposes that the 'predictor variables' are independent. Furthermore, logistic regression's performance may suffer when there is an imbalance in the data, when one class has more samples than another.

2) Random Forest

One prominent 'machine learning' method that excels at both regression and classification is random forest. It is classified as ensemble learning. It uses the combined predictive power of several ‘decision trees’ to increase prediction accuracy. The reality is that RF can handle high-dimensional data and does’nt need feature selection. or dimensionality reduction approaches is another reason for its popularity. Therefore, it frequently performs better than alternative algorithms in many circumstances. Additionally, Random Forest is resilient against missing data and outliers, which adds to its efficacy in many different data scenarios.

3) Decision Tree

A ‘decision tree’ is a prediction model in ' machine learning' that is visualized as a structure like a tree. The dataset is recursively divided into subsets according to characteristics, and choices are made at every node along the way, culminating in a leaf node that performs a prediction or classification. Because decision trees imitate people's decision-making processes, they are famous for being easy to understand and simple to use. They can offer details about the relevance of features and manage non-linear connections in data. Decision trees, however, may be sensitive to minute modifications in the dataset and are prone to excess fitting, particularly when presented with noisy data. Decision trees are frequently utilized in many different sectors because of their efficacy and simplicity of implementation, even with these drawbacks.

4) KNN

The K-Nearest Neighbors (KNN) algorithm is a popular and straightforward machine-learning technique that is useful for both regression and classification applications in water potability prediction tasks. KNN is a nucleotide method that doesn't make any approximations concerning the fundamental properties of the data, making it stand out. It's commonly known as a "lazy learner" algorithm because it uses iterations to progressively learn during the classroom instruction set. During the classification process, it applies the knowledge it has learned and stored instead of creating a model on the spot. By locating the KNN During the classroom instruction set, the 'KNN' technique establishes the class of a test sample. It finds the k closest neighbors, computes the distance between each 'training sample' and the 'testing sample', and gives the 'testing sample' the majority class label among its neighbors. 'KNN' is a straightforward yet powerful classification algorithm that is applied in many different fields because of its user-friendly design and simplicity of usage. The 'KNN' algorithms is a popular option for several 'classification and regression' applications because of its 'simplicity' and adaptability. Its simple methodology can handle both linear and nonlinear interactions in data and doesn't require a training phase. Nevertheless, when dataset sizes grow, KNN's computational complexity does as well.

5) SVM

When assessing whether or not samples of freshwater are suitable for human consumption, 'supervised machine learning algorithms' such as 'support vector machines (SVM)' are frequently employed. To determine this categorization, SVM looks into the "chemical and microbiological characteristics" of water samples. For big datasets including numerous dimensions and difficulties, there are two domains of nonlinear data evaluation in which this approach excels. Using the maximum margin idea, SVM builds a hyperplane that accurately classifies various water sample classes. An SVM model can accurately classify fresh samples as drinkable or non-drinkable by training on a labeled dataset of freshwater samples. Consequently, SVM is a crucial instrument for assessing the portability of freshwater and offers a dependable and precise way to guarantee the safety of drinking water.

IV. EXPERIMENTAL OUTCOMES

For the study, a dataset of 2,293 samples was utilized. Each sample underwent testing for there were nine separate water quality criteria, including pH, organic carbon, chloramines, turbidity, trihalomethanes, sulfate, hardness, conductivity, and solids. You can find a summary of these factors in the table. To simplify the analysis, The dataset was divided 80:20 between training and test data.

Table 2: Percentage of potable and non-potable water depending on the dataset

Potable	Non-Potable
39%	61%

In addition, the study looked at how well 'logistic regression', 'random forest', 'decision tree', KNN, and SVM could predict water potability. The study assessed the algorithms' performance using a confusion matrix and variables including 'true positive rate', 'false positive rate', 'true negative rate', and 'false negative rate'. The purpose of this extensive analysis was to provide informative information on the benefits and drawbacks of each algorithm for forecasting water quality.

Table 3: Proposed algorithms performance analysis

Model	Accuracy Score (%)	Precision	Recall	F1-Score
SVM	67.75	0.69	0.68	0.63
Random Forest	64.48	0.63	0.64	0.61
Decision Tree	63.61	0.62	0.64	0.58
Logistic Regression	61.87	0.38	0.62	0.47
KNN	61.43	0.60	0.61	0.60

It was clear based on the data presented in Table 3 that the 'SVM' model outperformed the other models. exhibiting outstanding outcomes. It obtained a precision of 0.69, an accuracy rate of 67.75%, and an F1-score of 0.63, indicating error-free classification performance.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

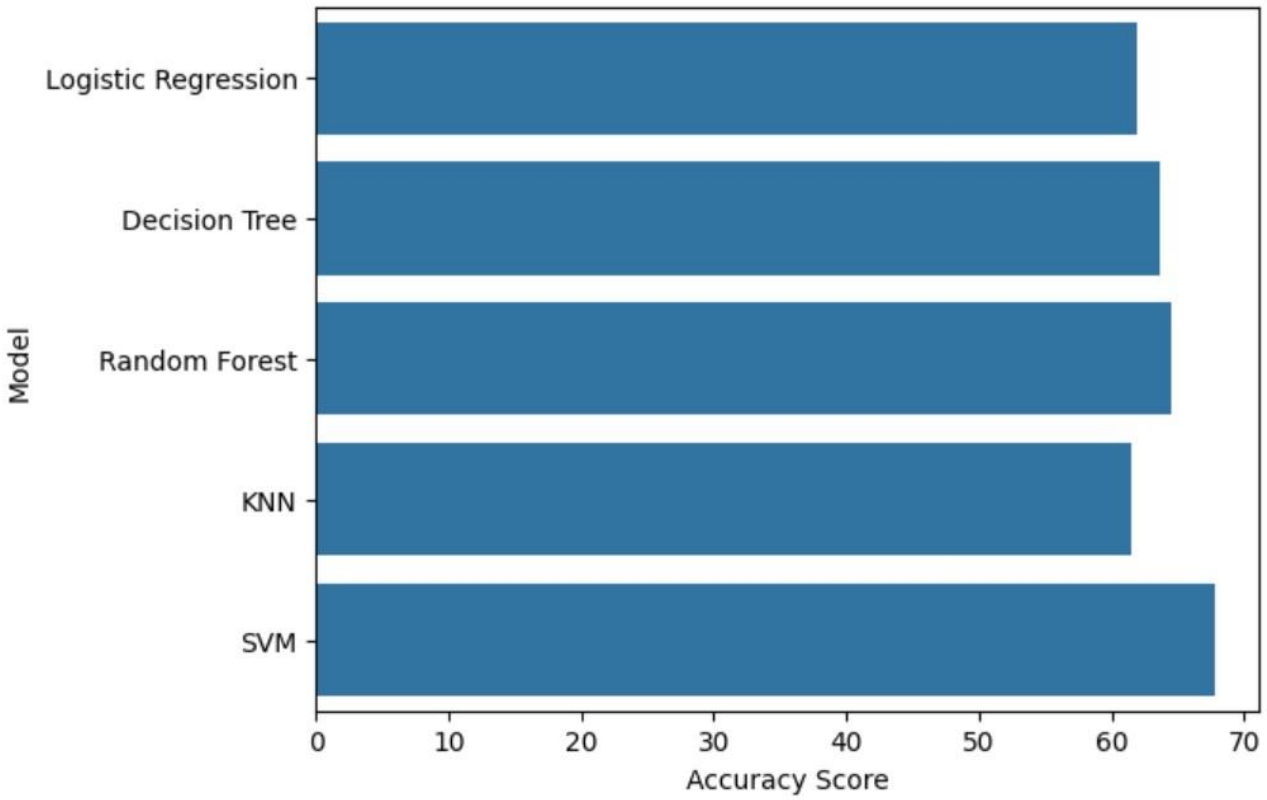


Fig 4: Accuracy comparison among the models

The study's accuracy scores for five ‘machine learning algorithms’ were displayed, as seen in Figure 4's bar graph. SVM has the best accuracy rate, depend on the graph, followed by ‘Random Forest’, ‘Decision Tree’, ‘Logistic Regression’, and ‘KNN’. By comparison, the method with the lowest accuracy rate was the ‘KNN’ model.

V. CONCLUSIONS

Ensuring that drinking water is safe and pure is essential for maintaining human health. Accurate forecasts of water potability are critical to achieving this aim. Everyone has the basic right to obtain clean drinking water because maintaining overall health and preventing waterborne infections are important. However, increasing global population and pollution levels have sparked grave concerns about the purity in water sources. Making use of machine-learning techniques might make a big difference in anticipating water potability and carrying out the necessary steps to enhance water quality, assuring that individuals have a chance to clean drinking water. It's crucial to recognize majority of the study's shortcomings, though. With just 2293 observations, the dataset employed in the study was fairly small. As such, extrapolating the findings to broader populations may be difficult. In addition, future research should consider additional pertinent aspects that may have an effect on the potability of water, as the current research concentrated on a restricted range of freshwater quality indicators.

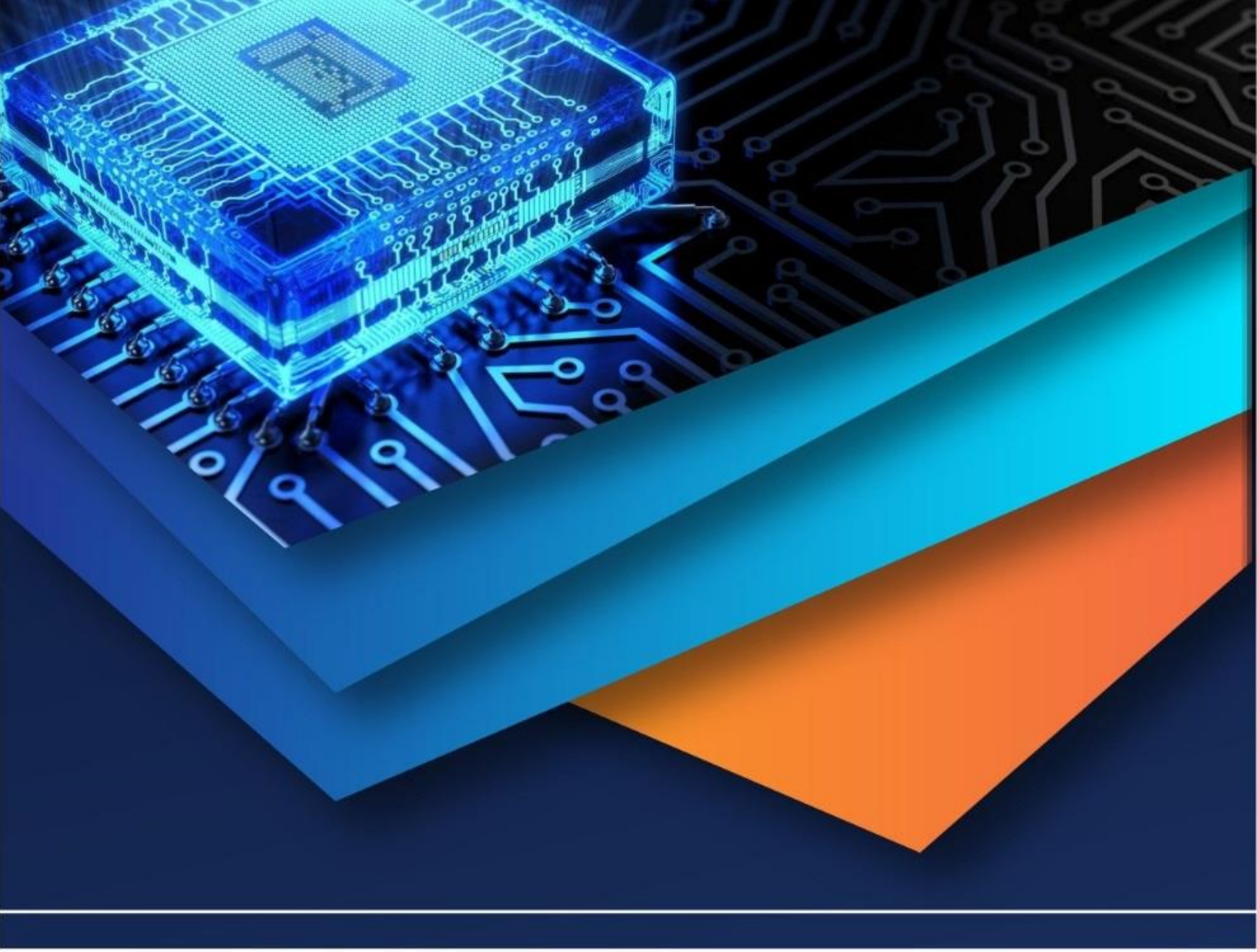
VI. ACKNOWLEDGMENT

Without acknowledging the people who made the project possible, without whose unwavering support and guidance my efforts would be considered vain., the happiness that comes with its successful completion would be incomplete. As we completed our project on "Predicting Potable Water Quality Using State-Of-Art ‘Machine Learning Algorithms’" we felt delighted to thank and show thanks to everyone who helped us along the way. We appreciate all of the help and inspiration we received along the way from our guide, Mr. Vijendra S. N., Assistant Professor of 'CSE'. For his assistance and direction, we are appreciative of Dr. Dhananjaya V., Professor and Head of 'CSE'. We are grateful to the management and principal of Impact College of Engineering and Applied Sciences, Dr. JALUMEDI BABU, for providing us with the facilities and welcoming atmosphere that have allowed us to further our education. We'd also want to mention all of the teaching and non-teaching staff of the 'CSE'. We'd want to mention. our parents and friends for their gracious cooperation and support over the project's duration.

REFERENCES

[1] Xin and Mou (2022). Multimodal-Based ML Algorithms for 'Water Quality' Classification: A Review. Wireless Communication and Mobile Computing in 2022.
[2] Alemayehu, D., Hackett, F., 2016. 'Water Quality' and Trophic State of Kaw Lake. 'Journal of Environmental' Studies 2, 1-7.
[3] https://www.kaggle.com/datasets
[4] https://washdata.org/
[5] "Potable Quality of Water Prediction Using ANN and ML algorithms" for Better Sustainability M Yurtsever, E Murat, 2023.
[6] Pal, O.K., "The Quality of Drinkable Water using 'ML' Techniques", 'Int. J. Adv. Eng'.
[7] Uddin, M.G., Nash, S., Rahman, A. and Olbert, "Performance analysis quality of the freshwater measure model for forecasting water state through ML techniques", Process Safety and Environmental Protection, 169, pp.808-828, 2023.
[8] Addisei M.B., "Evaluating Water Quality Parameters Used to Assess Drinking Water and Esthetic Attributes", Air, Soil and Water Research, 15, p.11786221221075005, 2022.
[9] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R. and García-Nieto, J., "Efficient quality of water prediction using 'supervised machine' learning", Water, 11(11), p.2210, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)