Trend Analysis of Water Physio-parameters of Mumbai, Pune, and Nagpur Using Machine Learning

SANTOSH SINGH[1], SIDDHESH MHATRE[2], VIJAY PRAJAPATI[3]

[1] *Head of Department, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*

[2 3] *PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*

Abstract— Rapid urbanization in cities like Mumbai, Pune, and Nagpur has engendered a myriad of challenges, chief among them being the preservation of water quality and sustainability. As urban centers burgeon, the demand for water resources escalates. However, this surge in demand is often juxtaposed with a deleterious rise in industrialization, infrastructural expansion, and the rampant discharge of sewage and pollutants into water bodies. These factors, in conjunction, have contributed to a progressive deterioration of water quality in urban areas, necessitating a paradigm shift in the management of this vital resource. The efficient management of water resources and the assurance of a clean drinking water supply have become pivotal issues confronted by municipal authorities. This research undertakes a comprehensive analysis of water quality trends in Mumbai, Pune, and Nagpur, employing advanced machine learning techniques to fathom complex datasets. These datasets encompass an array of physio-chemical parameters, such as pH, turbidity, electrical conductivity, and more, all of which are critical indicators of water pollution. The study requisitions historical water quality data procured from a multitude of sources, including rivers, lakes, and reservoirs. Subsequent preprocessing of this raw data precedes the application of machine learning algorithms like support vector machines, random forests, and k-nearest neighbors to unearth latent trends and patterns. The research stands upon a dual pedestal: first, it seeks to underscore the potential of machine learning as an indispensable tool for environmental monitoring and water quality analysis; second, it aspires to furnish urban planners and policymakers with actionable insights, paving the way for the mitigation of water pollution in these burgeoning cities. By aligning growth demands with environmental sustainability, the research aligns itself with the ultimate objective of fortifying the foundations of urban development with data-driven evidence, thereby orchestrating a harmonious balance between burgeoning urbanization and ecological preservation. In summary, this study endeavors to harness the power of data science to champion the cause of sustainable urban development, thereby ensuring the enduring sanctity of water quality in the cities of Mumbai, Pune, and Nagpur.

Indexed Terms— Trend Analysis, Supervised Learning Model, Classification

I. INTRODUCTION

In rapidly growing Indian cities like Mumbai, Pune, and Nagpur, ensuring a sustainable supply of clean and potable groundwater is of paramount importance. However, urbanization, industrialization, and increased infrastructure development have raised concerns about groundwater quality. The discharge of pollutants and contaminants into the ground has the potential to render groundwater unsuitable for both drinking purposes and supporting vital infrastructure. To address these challenges, our research takes a data-driven approach. We collect groundwater quality data from government websites, specifically focusing on key parameters that determine potability and infrastructure suitability. These parameters include pH levels, turbidity, dissolved oxygen, electrical conductivity, and more. Our analysis is based on historical data spanning several years, providing a comprehensive overview of the groundwater quality in these cities. By employing advanced machine learning algorithms, such as support vector machines, random forests, and neural networks, we aim to extract valuable insights and hidden trends from this dataset. The objectives of this research are twofold. Firstly, we aim to showcase the potential of leveraging machine learning and data analysis in environmental monitoring, especially in the context of groundwater quality assessment. Secondly, our research seeks to offer actionable insights to urban planners and decision-makers for effective groundwater management and environmental preservation in the face of rapid urban growth. Ultimately, this study endeavors to support sustainable urban development by using data-driven techniques to assess and manage groundwater quality. The outcomes are expected to inform evidence-based policies and decisions related to groundwater allocation, water source suitability, and infrastructure development, thereby ensuring the availability of safe drinking water and sustainable urban expansion.

II. MATERIALS AND METHODS

A. Study Area and Data Sources

The study focuses on analyzing water quality data from major urban centers in India - Mumbai, Pune, and Nagpur. Historical water quality data for these cities was obtained from government agency websites, including the Central Pollution Control Board (CPCB) portal. The dataset spans 5 years, from 2017 to 2021. It contains monthly measurements of various physico-chemical parameters from multiple sampling stations across rivers, lakes, and groundwater sources in each city.

B. Water Quality Parameters

The water quality parameters investigated include pH, turbidity, dissolved oxygen (DO), total dissolved solids (TDS), electrical conductivity (EC), alkalinity, hardness, chloride, fluoride, nitrate, sulfate, calcium, magnesium, sodium, and potassium. These parameters encompass the key indicators of water pollution, potability, and suitability for infrastructure projects.

C. Data Preprocessing

Feature Selection Correlation analysis and recursive feature elimination were utilized to select the most pertinent subset of features with maximal information content and minimal redundancy. This subset was used as input to machine learning models. Machine Learning Algorithms Supervised machine learning algorithms including SVM, random forest, and neural networks were explored for water quality prediction and classification tasks. The algorithms were implemented in Python using scikit-learn and Keras.

- Literature review

Theyazn H. H Aldhyani (2020) emphasizes the significance of modeling and predicting water quality due to the threat of various pollutants. The study focuses on developing advanced AI algorithms to predict the Water Quality Index (WQI) and classify water quality. The author employs artificial neural network models, particularly the nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM) deep learning algorithm, for WQI prediction. Additionally, three machine learning algorithms, support vector machine (SVM), k-nearest neighbor (K-NN), and Naive Bayes, are used for water quality classification (WQC) forecasting. The evaluation of the models using statistical parameters demonstrates their ability to accurately predict WQI and classify water quality. Specifically, the NARNET model performs slightly better than LSTM for WQI prediction, while the SVM algorithm achieves the highest accuracy of 97.01% for WQC prediction[1]. The paper published by Ozgur Kisi (2023) focuses on the estimation of water quality in the Yamuna River in Delhi, India, using hybrid neuro-fuzzy models. The study investigates the potential of four different neuro-fuzzy embedded meta-heuristic algorithms, namely particle swarm optimization, genetic algorithm, harmony search, and teaching–learning-based optimization algorithm, for accurate water quality prediction. The results indicate that using free ammonia, total Kjeldahl nitrogen, and water temperature as inputs improves the accuracy of COD prediction. The hybrid neuro-fuzzy models outperform the classical neuro-fuzzy model and LSSVM, achieving a 12% and 4% reduction in root mean square error, respectively [2]. Umair Ahmed's research (2023) focuses on using supervised machine learning to predict water quality. By utilizing four input parameters, such as temperature and turbidity, the study estimates the water quality index (WQI) and class (WQC). Gradient boosting and polynomial regression achieve efficient WQI prediction, with mean absolute errors (MAE) of 1.9642 and 2.7273 respectively. For WQC classification, the multi-layer perceptron achieves 85.07% accuracy. The research offers a cost-effective and faster approach for real-time water quality monitoring, contributing to efficient water resource management[3]. The study by Hongfang Lu (2020) introduces two hybrid decision tree-based machine learning models, CEEMDAN-RF and CEEMDAN-XGBoost, for accurate short-term water quality prediction. These models, incorporating the CEEMDAN data denoising technique, outperform conventional models in predicting water quality indicators. CEEMDAN-RF achieves low mean absolute percentage errors (MAPE) of 0.69% for temperature, 1.05% for dissolved oxygen, and 0.90% for specific conductance. CEEMDAN-XGBoost demonstrates MAPEs of 0.27% for pH, 14.94% for turbidity, and 1.59% for fluorescent dissolved organic matter. Overall, both models show superior performance with average MAPEs of 3.90% and 3.71% respectively, indicating their effectiveness in water quality prediction[4]. Dziri Jalal and Tahar Ezzedine's research addresses the critical issue of maintaining the quality of drinking water within distribution systems. They propose the use of wireless sensor networks to monitor and control water parameters, ensuring water quality adheres to established standards. Their innovative approach includes a real-time detection model based on machine learning to identify anomalies and potential malicious acts. To enhance efficiency, the research employs a data aggregation method to reduce processing time and data volume. Overall, their work aims to safeguard water quality and protect human and animal health in water distribution systems [5].

- Support Vector Machine

Support Vector Machine (SVM) stands as a cornerstone in the realm of machine learning algorithms, revered for its prowess in classification and regression tasks. At its core lies a sophisticated mechanism that seeks to identify a hyperplane, a multidimensional separator [4] that maximizes the distinction between data points of different classes while maintaining the widest possible gap, or margin, between these classes. SVM's versatility has granted it a venerated position in diverse domains, including image classification, text categorization, and even the intricate landscapes of bioinformatics. Its unique ability to navigate complex data distributions and high-dimensional spaces renders it indispensable in various data-driven endeavors. Working Principle: SVM's fundamental principle can be traced to its quest for a hyperplane that achieves a dual objective: first, it effectively segregates data points belonging to different classes, and second, it maximizes the distance between the hyperplane and the nearest data points of these classes. These closest data points are termed support vectors, pivotal entities that underpin SVM's decision-making process [8]. In essence, SVM strives to carve out a hyperplane that balances the act of classification accuracy and the imperative of generalization on unseen data. Linear SVM: In scenarios where data is linearly separable – amenable to separation through a straight line – a linear SVM unfurls its prowess. It engineers an optimal hyperplane, one that dissects the data classes while ensuring the maximal margin. This hyperplane is elegantly encapsulated in the equation `wx + b = 0`, where `w` symbolizes the weight vector and `b` represents the bias term. The weights, thus determined, guide SVM in classifying future data points. Non-linear SVM - Kernel Trick: However, real-world data often defies linearity, necessitating SVM to transcend its linear confines. This is where the kernel trick steps in – a wizardry that maps data to a higher-dimensional space where separability is achieved. The kernels, such as Polynomial, Gaussian (RBF), and Sigmoid, wield their transformative magic, enabling SVM to capture intricate relationships that could remain obscured otherwise. The kernel trick permits SVM to operate in realms where linear separation is a mirage. Parameters of SVM: The configurational levers of SVM amplify its adaptability and precision: - C (Cost Parameter): The parameter `C` governs the trade-off between widening the margin and tolerating misclassification. Smaller `C` values yield a larger margin but might admit misclassification, while larger `C` values constrain the margin to combat misclassification.

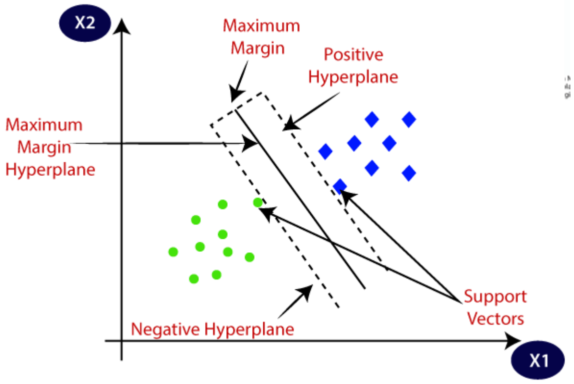

Fig. 1.0 SVM architecture

III. K-NEAREST NEIGHBORS

In the realm of machine learning and data analysis, K-Nearest Neighbors (KNN) stands out as a versatile, intuitive, and widely-used algorithm. KNN is part of the supervised learning paradigm, primarily applied for classification and regression tasks. It's a fundamental algorithm that, despite its simplicity, plays a vital role in various real-world applications. In this essay, we will delve into the K-Nearest Neighbors algorithm, exploring its underlying principles, applications, advantages, limitations, and the critical factors that affect its performance.

The Core Concept of K-Nearest Neighbors At its heart, K-Nearest Neighbors is a non-parametric, instance-based learning algorithm. This means that it doesn't make any assumptions about the data distribution, unlike many other algorithms that are based on statistical models. Instead, KNN classifies or predicts based on the similarity between data points in a feature space. The "K" in KNN represents the number of nearest neighbors to consider when making predictions. It's a hyper parameter that needs to be specified before applying the algorithm. To understand KNN better, let's break down its core concepts: Data Collection: In any KNN application, the first step involves gathering a dataset containing labeled examples. Each example consists of a set of features and their corresponding class or value. For example, in a classification task, the dataset might contain features of various animals and their respective classes, such as "dog," "cat," "bird," etc. K-Value: One of the critical decisions to make when using KNN is to choose an appropriate value for K. This determines the number of neighbors that will be considered when making predictions. Selecting the right K-value can significantly impact the algorithm's performance. A small K may result in a noisy prediction, while a large K may cause over-smoothing of predictions. Distance Metric: KNN relies on a distance metric to measure the similarity between data points. The choice of distance metric can significantly influence the algorithm's performance. Common distance metrics include Euclidean distance, Manhattan distance, Murkowski distance, and others. The selection of the metric should be guided by the nature of the data and the problem being solved.

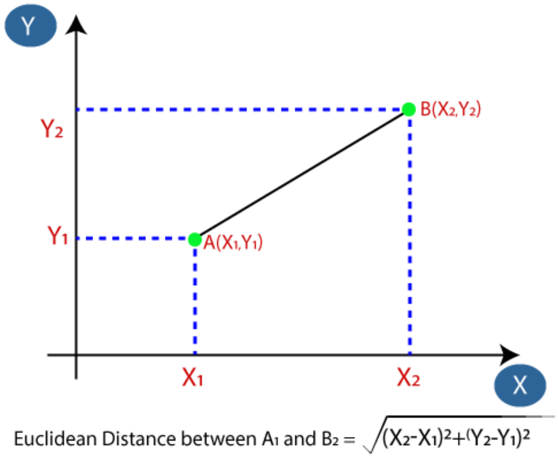

Fig. 1.1 KNN architecture

IV. METHODOLOGY

The first step is to accumulate reliable and accurate water quality data from an reliable authorities website. This facts will then undergo a statistics transformation procedure to handle any mistakes or inconsistencies and make sure it's miles in a appropriate format for analysis. Next, exploratory facts analysis (EDA) strategies may be hired to advantage a deeper knowledge of the records. This entails visualizing the statistics, identifying patterns, and analyzing relationships among distinct variables. EDA allows in uncovering insights and ability correlations that may make contributions to predicting water satisfactory. After acting EDA, the data might be split into schooling and checking out sets. The schooling set can be used to teach the gadget mastering version, allowing it to study patterns and relationships from the data. The checking out set might be used to evaluate the overall performance of the educated version, imparting an estimate of its predictive accuracy. Finally, a suitable machine studying set of rules could be chosen primarily based on the character of the trouble and the available data. The decided on algorithm can be implemented to the training set to create a predictive version [6]. This version will then be used to make predictions at the checking out set, supplying precious insights into the destiny water high-quality primarily based on the learned styles and relationships from the education facts. Throughout the entire method, proper records handling techniques, feature choice, and model assessment could be done to make certain the accuracy and reliability of the water excellent predictions. In addition to those steps, it is vital to note that system gaining knowledge of models are not perfect and might not continually offer accurate predictions. Therefore, it's far essential to constantly monitor and update the model with new data to improve its overall performance through the years. Furthermore, it is critical to don't forget outside factors that may affect water first-class, together with climate conditions, human sports, and herbal screw ups. These elements should be taken under consideration when making predictions and decoding the consequences. Overall, predicting water exceptional the use of machine studying involves a mixture of records series, preprocessing, evaluation, modeling, and assessment [7]. By following this system, it's far possible to develop a dependable and accurate predictive version for water exceptional. One of the important thing factors of this system is the selection of a appropriate machine learning set of rules. There are many one of a kind algorithms to pick out from, every with its own strengths and weaknesses. Some commonplace algorithms used for predictive modeling encompass selection timber, random forests, guide vector machines, and neural networks. The choice of set of rules will depend on the character of the hassle and the to be had data. Working Principle: SVM's essential principle can be traced to its quest for a hyperplane that achieves a twin goal: first, it efficiently segregates information factors belonging to one of a kind instructions, and 2d, it maximizes the distance between the hyperplane and the closest information factors of those lessons. These closest statistics points are termed assist vectors, pivotal entities that underpin SVM's choice-making process. In essence, SVM strives to carve out a hyperplane that balances the act of type accuracy and the vital of generalization on unseen information.

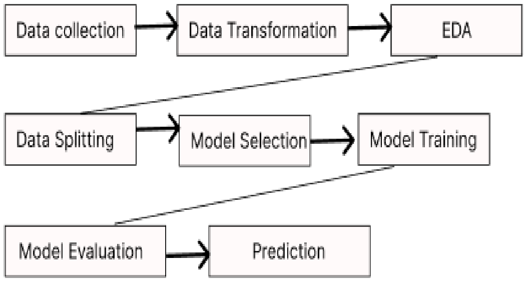

Fig. 0.0.1 Methodology

The workflow of the KNN algorithm is straightforward:
Data Collection: Gather the dataset with labeled examples. Choose K: Select a value for K, which determines the number of neighbors to consider during prediction. Calculate Distances: For a given data point to be classified or predicted, calculate the distances to all other data points in the dataset using the chosen distance metric. Identify Neighbors: Select the K data points with the shortest distances to the target data point. These data points are the "nearest neighbors." Majority Vote (Classification) or Averaging (Regression): For classification tasks, count the occurrences of each class among the K neighbors and assign the class with the highest count to the target data point. In regression tasks, compute the average of the values associated with the K neighbors. The KNN algorithm is robust, easy to understand, and suitable for both classification and regression tasks. It can handle multi-class problems effectively, and its outcomes can be interpreted intuitively. KNN offers several advantages: Simplicity and Ease of Implementation: KNN's simplicity and lack of complex model training make it a go-to choice, especially when you need to prototype quickly or have limited labeled data. No Assumptions about Data Distribution: Unlike parametric models like linear regression, KNN doesn't make assumptions about the underlying data distribution. It can work well with data that doesn't conform to any specific mathematical model.Versatility: KNN is suitable for both classification and regression tasks. It can handle a wide range of problems, from predicting house prices based on features to classifying images of objects or animals. Multi-Class Problems: KNN naturally extends to multi-class classification problems, making it applicable in scenarios where there are more than two possible classes. Intuitive Interpretation: KNN's predictions are easy to interpret. In classification, a data point is assigned to the class that is most frequent among its K-nearest neighbors. In regression, the predicted value is simply the average of the values of those neighbors. Non-Linear Decision Boundaries: KNN can capture complex, non-linear decision boundaries in the data, making it suitable for problems where linear models may not perform well. Ease of Updating: KNN is amenable to online learning, allowing for incremental updates to the model as new data becomes available.

V. RESULTS

In this section, we present the partial results of applying the Support Vector Machine (SVM) algorithm to two different tasks "Potability" and "Infrastructure_Suitability" prediction. We explored the SVM classifier's performance for each task, and the evaluation includes the variation of the regularization parameter (C) to fine-tune the models.Potability Prediction For the "Potability" prediction task, we selected a set of water quality features from our dataset, including 'Ca', 'Cl', 'CO3', 'EC', 'HCO3', 'K', 'Mg', 'Na', 'NO3', 'pH', 'SO4', 'TDS', 'TH', and 'F'. After splitting the data into training and testing sets, we applied SVM classifiers with varying values of the regularization parameter (C) ranging from 1 to 9. The results indicate that, by iterating over different C values, we achieved the highest accuracy of approximately 93% for predicting "Potability" when C was set to 9. The SVM model demonstrates the potential to effectively classify water samples into potable and non-potable categories. The training and testing accuracy curves showcase the model's performance across different levels of regularization, with testing accuracy steadily increasing as C is adjusted.

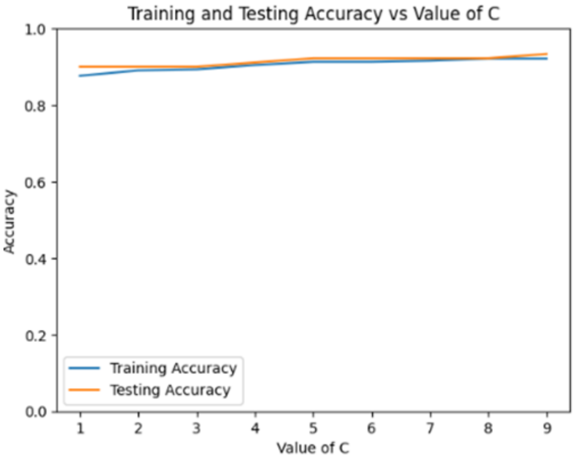

Fig. 2.0 Accuracy graph for Potability

In the "Infrastructure_Suitability" prediction task, a distinct set of water quality features was chosen, including 'CO3', 'EC', 'K', 'Mg', 'Na', 'NO3', 'pH', 'SO4', 'TDS', 'TH', and 'F'. Similar to the "Potability" task, SVM classifiers were employed with different C values to optimize model performance.

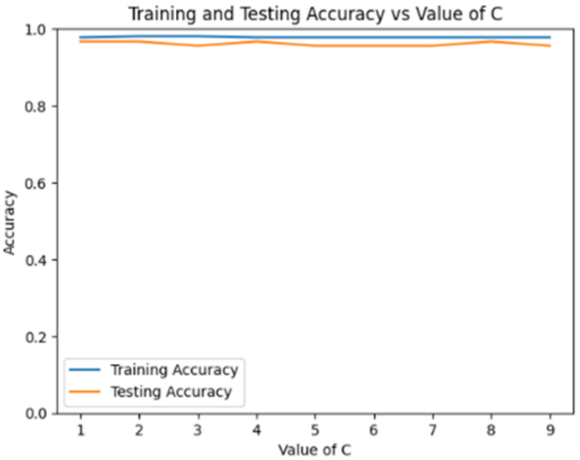

Fig. 2.1 Accuracy graph for Infrastructure_Suitability

The SVM model yielded impressive results for predicting "Infrastructure_Suitability." With C set to 3, the model achieved a remarkable testing accuracy of approximately 95%, demonstrating the ability to categorize water samples effectively based on infrastructure suitability. The training and testing accuracy curves reveal a consistent performance improvement as the regularization parameter C is adjusted, indicating that the model adapts to the data's characteristics.

In this section, we present the partial results of applying the K-Nearest Neighbors (KNN) algorithm to two different tasks within our dataset. These tasks involve predicting "Potability" and "Infrastructure_Suitability" based on various water quality features. We have utilized the KNeighborsClassifier from scikit-learn for classification and have evaluated the model's performance by computing accuracy and visualizing the confusion matrix. Potability Prediction For the "Potability" prediction task, we selected a subset of features from our dataset, including 'Ca', 'Cl', 'CO3', 'EC', 'HCO3', 'K', 'Mg', 'Na', 'NO3', 'pH', 'SO4', 'TDS', 'TH', and 'F'. After splitting the data into training and testing sets, we trained a KNN classifier with five neighbors and made predictions on the test set.The accuracy of the KNN model for predicting "Potability" is approximately 94%, indicating that the model's performance is highly promising.

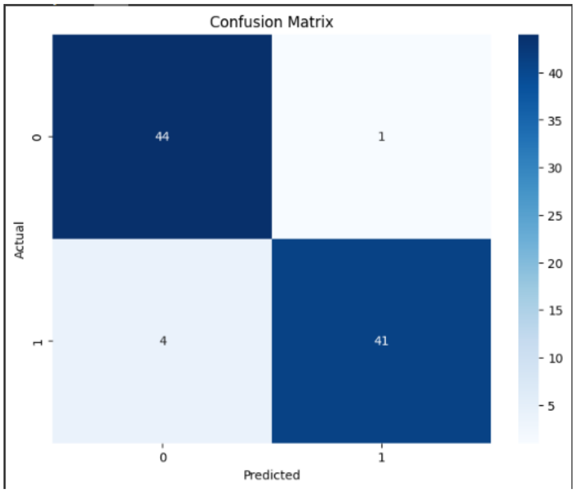

Fig. 2..2 Confusion matrix Infrastructure_Suitability

The confusion matrix visualization shows that the model has successfully classified instances into their respective classes. Infrastructure Suitability Prediction In the "Infrastructure_Suitability" prediction task, we selected a different set of features, including 'CO3', 'EC', 'K', 'Mg', 'Na', 'NO3', 'pH', 'SO4', 'TDS', 'TH', and 'F'. Similarly, we trained a KNN classifier with five neighbors and evaluated its performance.Remarkably, the KNN model achieved an accuracy of approximately 94% for predicting "Infrastructure_Suitability." This suggests that the model can effectively categorize data points into different infrastructure suitability categories. The confusion matrix visualization further demonstrates the model's capability to make accurate predictions.These partial results showcase the potential of the KNN algorithm for both classification tasks within our dataset. The high accuracy achieved in both cases suggests that KNN may be a suitable choice for these specific prediction tasks. However, it's essential to keep in mind that the ultimate success of a machine learning model depends on the nature of the data and the specific problem context.

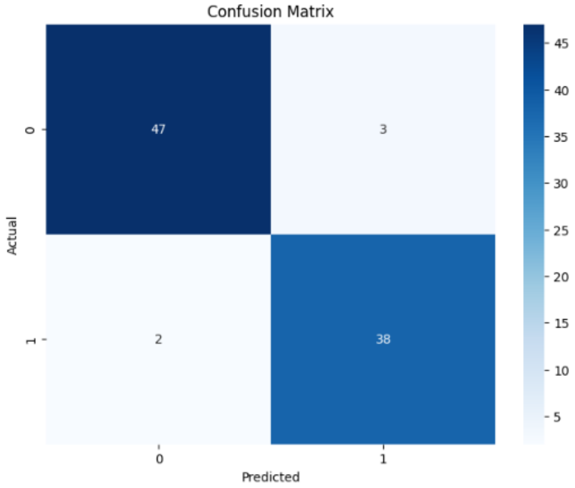

Fig 3.0 Potability confusion matrix

CONCLUSION

The research also demonstrates the efficacy of machine learning in water quality analysis. The algorithms proved adept at detecting subtle changes and linkages between parameters compared to conventional statistical approaches. This highlights the merit of integrating data science capabilities into environmental monitoring and urban planning. However, fully realizing the potential of these findings requires active collaboration between scientists, policymakers, and urban stakeholders. Synthesizing the data-driven insights with on-ground domain expertise and socio-economic considerations is key to developing holistic solutions. This interdisciplinary approach is imperative for balancing the demands of development with sustainability.

While the current models provide a robust foundation, enhancements in predictive accuracy, model interpretability, and integration of socio-economic factors could enrich future iterations. As cities continue to evolve, maintaining clean and safe water will require continuous innovation. In conclusion, this trend analysis serves as an important step towards evidence-based management of water resources in Mumbai, Pune, and Nagpur. The principles and methods established can inform sustainable urban planning across India and beyond. But fully unlocking the value of data-driven insights requires an integrated effort between science, governance, and society.

REFERENCES

[1] Theyazn H. H Aldhyani (2020) - "Advanced AI Algorithms for Predicting Water Quality and Classification Using Machine Learning"

[2] Kisi, O., Parmar, K.S., Mahdavi-Meymand, A., Adnan, R.M., Shahid, S. and Zounemat-Kermani, M., 2023. Water quality prediction of the yamuna river in India using hybrid neuro-fuzzy models. Water, 15(6), p.1095.

[3] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R. and García-Nieto, J., 2019. Efficient water quality prediction using supervised machine learning. Water, 11(11), p.2210.

[4] Lu, H. and Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, p.126169.

[5] Adu-Manu, K.S., Tapparello, C., Heinzelman, W., Katsriku, F.A. and Abdulai, J.D., 2017. Water quality monitoring using wireless sensor networks: Current trends and future research directions. ACM Transactions on Sensor Networks (TOSN), 13(1), pp.1-41.

[6] Khullar, S. and Singh, N., 2021. Machine learning techniques in river water quality modelling: a research travelogue. Water Supply, 21(1), pp.1-13.

[7] Suthar, S., Bishnoi, P., Singh, S., Mutiyar, P.K., Nema, A.K. and Patil, N.S., 2009. Nitrate contamination in groundwater of some rural areas of Rajasthan, India. Journal of hazardous materials, 171(1-3), pp.189-199.

[8] Patil, P.N., Sawant, D.V. and Deshmukh, R.N., 2012. Physico-chemical parameters for testing of water-a review. International journal of environmental sciences, 3(3), p.1194.

[9] Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A. and Al-Shamma'a, A., 2022. Water quality classification using machine learning algorithms. Journal of Water Process Engineering, 48, p.102920.

[10] Rehana, S. and Mujumdar, P.P., 2012. Climate change induced risk in water quality control problems. Journal of Hydrology, 444, pp.63-77.

[11] Juahir, H., Zain, S.M., Yusoff, M.K., Hanidza, T.T., Armi, A.M., Toriman, M.E. and Mokhtar, M., 2011. Spatial water quality assessment of Langat River Basin (Malaysia)