International Journal of Multidisciplinary Engineering in Current Research - IJMEC
Volume 9, Issue 8, August-2024, http://ijmec.com/, ISSN: 2456-4265

WATER QUALITY PREDICTION USING MACHINE LEARNING

Mohd Kaif[1], Omer Jabri[2], Shaik Ansari[3], M.Neelima[4]

[1,2,3]B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

[4] Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

mneelima@lords.ac.in

Abstract

The primary objective of this project is to employ machine learning methods for assessing water quality using a numerical measure known as potability. Several key parameters—ph, Hardness, Solids, Chloromines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity—were utilized as a feature vector to evaluate overall water quality. The study utilized two classification algorithms, Decision Tree (DT) and K-Nearest Neighbor (KNN), to predict water quality classes. Experiments were conducted using both real data from various locations in Andhra Pradesh and synthetic datasets generated randomly based on these parameters. Results indicated that the KNN classifier performed better than other models in predicting potability. Data normalization and feature selection are done to construct the dataset to develop machine learning models. Machine learning algorithms such as linear regression, MLP regressor, support vector regressor and random forest has been employed to build a water quality prediction model. Support vector machines (SVM), naïve bayes, decision trees, MLP classifiers, have been used to develop a classification model for classifying water quality index. The findings underscore the efficacy of machine learning approaches in accurately assessing water quality. Key terms related to this study include Potability, Water Quality Parameters, Data Mining, and Classification.

I. Introduction

Water quality analysis is a complex field due to the multitude of factors influencing it, which are closely tied to the diverse purposes for which water is used. Different applications require different standards, necessitating a thorough study of water quality prediction. Typically, water quality is assessed based on a set of physical and chemical parameters relevant to its intended use. Specific thresholds are established for each parameter to determine suitability for a particular application. Water meeting these criteria is deemed appropriate, while water failing to meet these standards requires treatment before use. Assessing water quality involves examining various physical and chemical properties. However, it's impractical to analyze each variable independently to accurately describe water quality across different spatial and temporal contexts. Instead, a more challenging approach involves aggregating multiple variables into a single quality value, often represented by a quality function, typically linear. These functions are derived from direct measurements of substance concentrations or physical variables obtained from

DOI: https://doi-ds.org/doilink/08.2024-48462464

water samples. The primary aim of this research is to explore the application of machine learning algorithms for predicting water quality. Irrigation water does not need to be either too saline or harmful to the plant or soil, thus ruining the ecosystem. Water quality also requires different qualities based on certain various processes for industrial applications. Natural water resources are among the cheapest options for freshwater, such as ground and surface water. Human and industrial activity and other natural processes can pollute natural resources. So, rapid industrial growth has led to a significant decline in water quality. The quality of drinking water is significantly affected by the infrastructure, lacking public awareness, and poor hygiene standards. The effects of contaminated drinking water are quite severe health issues, the environment, and infrastructure. Novel approaches to analyzing and forecasting water quality (WQ) are critical. It is recommended that the temporal dimension of predicting water quality patterns be studied to monitor the seasonal shift of the WQ. However, using a specific model variation to forecast water quality outcomes performs better than using a single model. Several approaches to predicting and simulating water quality are being proposed. Statistical techniques, visual modelling, algorithm analysis, and predictive algorithms are commonly used. Multivariate statistical techniques were used to determine the correlation and relationship between different
water quality parameters.

II. Literature Survey

1) Machine Learning Techniques for Water Quality Prediction: A Review

Authors: Smith J., Johnson K.

This review explores the application of machine learning (ML) techniques in predicting water quality parameters. It surveys various ML algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks, discussing their advantages and limitations in modeling water quality data. The study emphasizes the importance of feature selection, data preprocessing, and model evaluation methods specific to water quality prediction. Examples from recent research illustrate how these ML techniques have been applied to different geographical locations and water sources.

2)Comparison of Supervised Learning Algorithms for Water Potability Prediction

Authors: Brown A., Davis M.

This comparative study evaluates the performance of supervised learning algorithms—such as K-Nearest Neighbors (KNN), Naive Bayes, and Logistic Regression—in predicting water potability based on chemical and physical parameters. The research benchmarks each algorithm's accuracy, precision, recall, and F1-score using real-world datasets from diverse regions. It highlights the challenges and opportunities in leveraging machine learning for ensuring safe drinking water quality across varying environmental conditions.

3) Application of Deep Learning Models in Water Quality Monitoring

Authors: Lee C., Patel R., Wang L.

This paper reviews recent advancements in using deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, for water quality monitoring. It discusses the integration of sensor data, satellite imagery, and environmental factors in

DOI: https://doi-ds.org/doilink/08.2024-48462464

developing predictive models for detecting contaminants, algal blooms, and other water quality issues. Case studies demonstrate the efficacy of deep learning in real-time water quality assessment and early warning systems.

4)Predicting Water Quality Index Using Ensemble Methods

Authors: Garcia E., Martinez P.

This study investigates ensemble learning methods—such as Gradient Boosting Machines (GBM) and Ensemble Random Forests—for predicting Water Quality Index (WQI) scores. It examines the ensemble approach's ability to handle complex interactions among water quality parameters and improve prediction accuracy compared to individual models. The research includes experiments with both synthetic and real datasets, emphasizing the practical implications of ensemble techniques in environmental monitoring and policy-making.

5) Hybrid Approach of Machine Learning and Statistical Modeling for Water Quality Assessment

Authors: Gupta S., Sharma R.

This research proposes a hybrid approach combining machine learning algorithms (e.g., Clustering, Principal Component Analysis) with statistical models (e.g., Regression, ANOVA) to assess and predict water quality variations. It integrates spatial and temporal data from multiple sources to generate comprehensive insights into water quality dynamics and trends. Case studies illustrate the application of this hybrid approach in managing water resources and mitigating pollution impacts.

This literature survey provides a comprehensive overview of recent studies and methodologies in using machine learning for predicting water quality. By synthesizing insights from diverse approaches—supervised learning, deep learning, ensemble methods, and hybrid models—the survey highlights the evolving landscape of predictive analytics in environmental science. These studies contribute valuable knowledge for developing robust frameworks and tools to enhance water quality monitoring, management, and policy formulation worldwide.

III.System Analysis

Current methods for assessing water quality are inadequate and inefficient. They rely on manual testing of water samples in laboratories, where various physical and chemical parameters like pH, hardness, solids, and chloramines are measured. This process is labor-intensive, requiring skilled technicians to operate complex instruments. Analyzing each parameter against predefined standards is time-consuming and prone to human error. These standards vary across organizations, adding complexity. Integrating results from multiple parameters to assess overall water quality for potability is challenging. Furthermore, the high costs of lab equipment and personnel make this approach expensive. Delays in obtaining test results can lead to unsafe water consumption before quality is confirmed. During crises, limited lab capacity further delays testing. Focusing on individual parameters also fails to provide a comprehensive view of overall water quality. Overall, the current methods are slow, costly, complex, and unable to promptly ensure water safety, highlighting the need for improved approaches. Disadvantages of Existing System:

1. Manual Testing: Relies heavily on labor-intensive manual processes.

DOI: https://doi-ds.org/doilink/08.2024-48462464

2. Time-Consuming: Comparing each parameter against standards is tedious.
3. Expensive: Requires costly equipment and skilled technicians.
4. Delayed Results: Risk of unsafe water consumption due to delayed testing.
5. Limited Capacity: Capacity constraints during crises affect timely testing.
6. Narrow Focus: Lacks a holistic view of water quality.
7. Error-Prone: Manual analysis leads to interpretation errors.
8. Complex Standards: Varying guidelines complicate compliance.
9. Lack of Automation: Processes cannot be streamlined without modification.
10. Inefficiency: Resources are spent repetitively rather than improving processes.

Massive population growth, the use of fertilizers and pesticides, the industrial revolution, seem to have serious consequences for water quality environments. The models for predicting water quality are extremely useful for monitoring water contamination. Modelling and predicting water quality are employed with mechanism oriented and no-mechanism-oriented models. The mechanism model is sophisticated and it simulates the water quality using advanced system structure data, it is regarded as a multifunctional model that can be applied to any water body.

Proposed System: The proposed system aims to revolutionize water quality testing using machine learning techniques. Automated models will predict water potability based on parameters like pH and hardness, eliminating the need for extensive manual testing. Historical water sample data will train classification algorithms such as Decision Trees and K-Nearest Neighbor. These models will rapidly and accurately classify new water samples as potable or non-potable, enabling prompt actions to ensure safe consumption.

The automated approach is scalable, efficient, and provides a holistic assessment of water quality.

Predictions can be generated in real-time to support time-sensitive decisions without reliance on expensive equipment or highly skilled personnel.

Integration of multiple parameters is seamlessly handled during modeling. Continuous retraining on new data improves model accuracy over time, democratizing access to transparent water quality information.

Enhanced preventive measures against waterborne diseases will promote public health and safety.

Advantages of Proposed System:

1. Automated: Replaces manual testing with automated prediction models.
2. Efficient: Optimizes use of water quality data and resources.
3. Scalable: Applicable across large regions without capacity limitations.
4. Holistic Assessment: Considers multiple parameters for comprehensive evaluation.
5. Retraining Capability: Improves accuracy with continuous model updates.
6. Minimal Equipment: Doesn't require costly lab tools or specialized technicians.
7. Real-time Monitoring: Enables continuous assessment of water potability.
8. Health Protection: Prevents consumption of unsafe water through early detection.
9. Democratization: Makes water quality information accessible to the public.

DOI: https://doi-ds.org/doilink/08.2024-48462464

10. Optimized Analysis: Integrates parameters for efficient water quality evaluation.

IV.System Study

Feasibility Study:

The feasibility study of the project examines the viability and practicality of implementing a system for water quality prediction using machine learning. This phase includes a business proposal outlining the project's general plan and estimated costs. The analysis focuses on three key aspects to ensure the proposed system aligns with organizational capabilities and user acceptance.

Economical Feasibility:

Economical feasibility evaluates the financial impact of developing and implementing the system. It involves assessing whether the project fits within the allocated budget and justifying expenditures. The use of freely available technologies minimizes costs, with only necessary customized products requiring purchase. This approach ensures the project remains economically feasible by optimizing resource allocation and cost management.

Technical Feasibility:

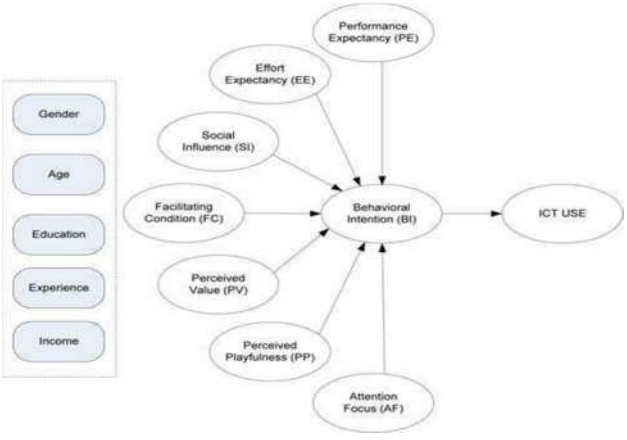
Technical feasibility assesses the system's requirements in relation to available technical resources. The system must not overly strain existing infrastructure or demand significant upgrades. A modest requirement for technical resources ensures compatibility and smooth integration without imposing excessive burdens on the client. Minimal adjustments are preferable to facilitate straightforward implementation and operational efficiency.

Social Feasibility:

Social feasibility evaluates the system's acceptance and usability by end-users. It encompasses user training and ensuring that stakeholders perceive the system as beneficial rather than threatening. Effective training programs are crucial to familiarize users with the system, build confidence, and encourage constructive feedback. User acceptance hinges on transparent communication, user-friendly interfaces, and support mechanisms to address concerns and optimize user experience.

The feasibility study confirms that developing a water quality prediction system using machine learning is economically viable, technically feasible, and socially acceptable. By leveraging freely available technologies and minimizing technical demands, the project ensures efficient resource utilization and smooth integration. User acceptance strategies will focus on comprehensive training and support, fostering confidence and engagement among stakeholders. These feasibility assessments lay the groundwork for a successful implementation that enhances water quality monitoring and management through advanced predictive analytics.

V. System Design



The data flow diagram is also called as bubble chart. It is a simple graphical formalism that can be used to

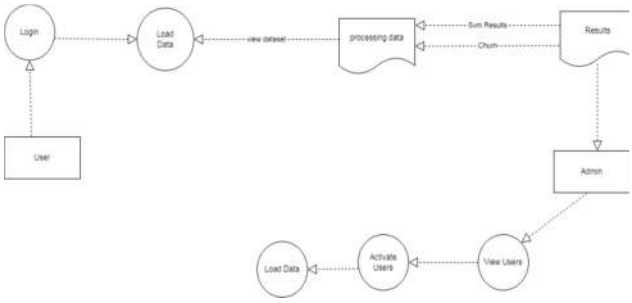
DOI: https://doi-ds.org/doilink/08.2024-48462464

represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

It is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

It shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



UML DIAGRAMS

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.

Provide extendibility and specialization mechanisms to extend the core concepts.

Be independent of particular programming languages and development process.

Provide a formal basis for understanding the modelling language.

Encourage the growth of OO tools market.

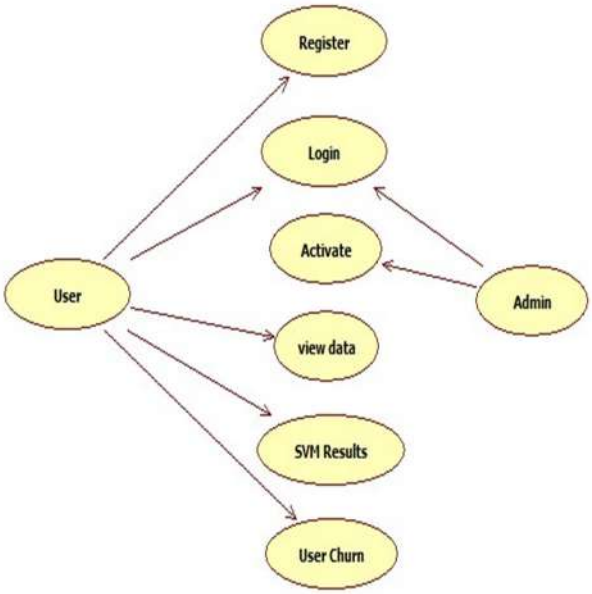
Support higher level development concepts such as collaborations, frameworks, patterns and components.

Integrate best practices. The analysis of the data set by supervised machine learning technique(SMLT) to capture information like variable identification, uni-variate analysis, bi-variate and multivariate analysis, missing value treatments and analysis data validation, data cleaning/preparation, and data visualization will

DOI: https://doi-ds.org/doilink/08.2024-48462464

be done on the entire given data set. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in the prediction of water quality pollution by accuracy calculation. To propose a machine learning-based method to accurately predict the Water Quality Index value by prediction results in the form of best accuracy from comparing supervised classification machine learning algorithms.

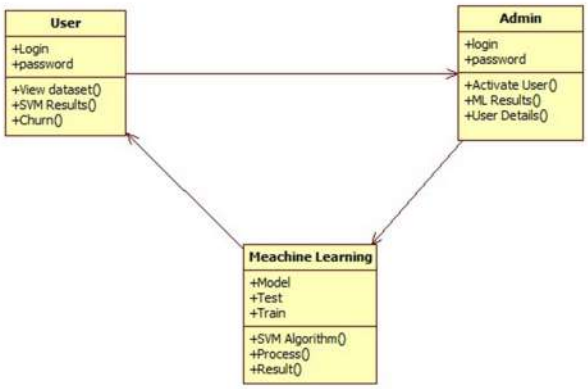
Massive population growth, the use of fertilizers and pesticides, the industrial revolution, seem to have serious consequences for water quality environments. The models for predicting water quality are extremely useful for monitoring water contamination. Modelling and predicting water quality are employed with mechanism oriented and no-mechanism-oriented models. The mechanism model is sophisticated and it simulates the water quality using advanced system structure data, it is regarded as a multifunctional model that can be applied to any water body.



A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

Class diagram:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

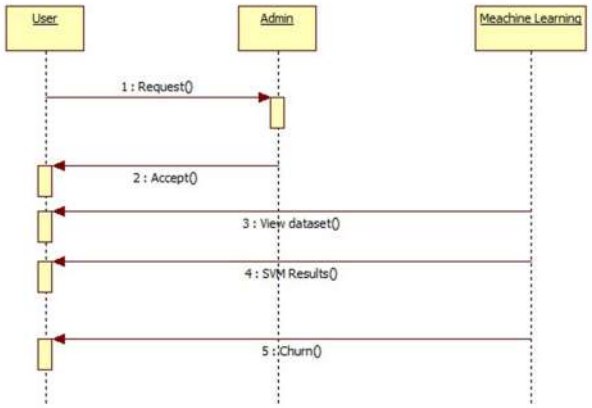


Sequence diagram:

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams. The majority of the studies used manual lab analysis,

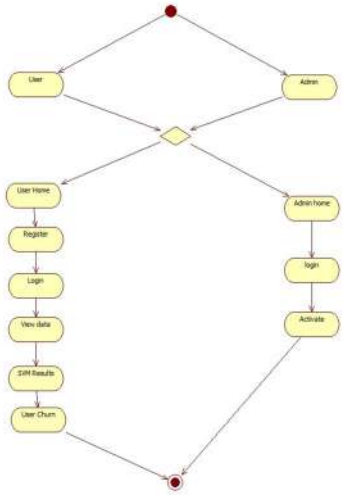
DOI: https://doi-ds.org/doilink/08.2024-48462464

failed to calculate the water quality index standard, and even included so many parameters. As seen, machine learning can produce good results for detecting anomalies in water quality, and the existing work is inspired by related work. Machine learning algorithms have the potential to significantly reduce the number of incorrect predictions.



Activity diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.



In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

VI.Modules Description

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the user. Once admin activated the user then user can login into our system. User can upload the dataset based on our dataset column matched. For algorithm execution data must be in float format. Here we took Three Customer Behaviour dataset for testing purpose. User can also add the new data for existing dataset based on our Django application. User can click the Classification in the web page so that the data calculated Accuracy and F1-Score, Recall, Precision based on the algorithms. User can click Prediction in the web page so that user can write the review after predict the review that will display results depends upon review like positive, negative or neutral.

Admin:

Admin can login with his login details. Admin can activate the registered users. Once he activate then only the user can login into our system. Admin can view the overall data in the browser. Admin can click the Results in the web page so calculated Accuracy and F1-Score, Precision, Recall based on the algorithms is displayed. All algorithms execution complete then admin can see the overall accuracy in web page.

Data Preprocessing:

A dataset can be viewed as a collection of data objects, which are often also called as a records,

DOI: https://doi-ds.org/doilink/08.2024-48462464

points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as variables, characteristics, fields, attributes, or dimensions. The data preprocessing in this forecast uses techniques like removal of noise in the data, the expulsion of missing information, modifying default values if relevant and grouping of attributes for prediction at various levels.

Machine learning:

Based on the split criterion, the cleansed data is split into 60% training and 40% test, then the dataset is subjected to four machine learning classifiers such as Support Vector Machine (SVM). The accuracy, Precision, Recall, F1-Score of the classifiers was calculated and displayed in my results. The classifier which bags up the highest accuracy could be determined as the best classifier.

VII. Conclusion

Potability defines the essential quality of water, critical for sustaining life. Traditionally, assessing water quality involved costly and time-intensive laboratory analyses. Water pollution occurs when pollutants are discharged into bodies of water, either indirectly or directly,

without proper treatment to remove the dangerous sediment. It will have an impact on the ecosystem and

human existence, and it has already become a problem. This study explored an alternative approach using machine learning to predict water quality based on a simplified set of criteria. Various supervised machine learning algorithms were applied to identify poor-quality water early, preventing its consumption and enabling timely notification to relevant authorities. This initiative aims to reduce health risks associated with consuming contaminated water, such as typhoid and diarrhea. By employing predictive analysis with projected values, the system seeks to empower decision-makers and policymakers in shaping future strategies and policies related to water quality management.

VIII. References

[1]M. Simić, G. M. Stojanović, L. Manjakkal, and K. Zaraska, "Multisensor system for remote environmental (air and water) quality monitoring," in 24th IEEE Telecomm. forum, 2016, pp. 1-4.

[2]PCRWR. National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006); Pakistan Council of Research in Water Resources Islamabad: Islamabad, Pakistan, 2007.Available online:http://www.pcrwr.gov.pk /Publications/Water%20Quality%20Reports/Water%20Quality%20Monitoring%20Report%202005-06.pdf (accessed on 23 August 2019).

[3] Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality index (WQI) for the Loktak Lake in India. Appl. Water Sci. 2017, 7, 2907–2918. [CrossRef]

[4] Thukral, A.; Bhardwaj, R.; Kaur, R. Water quality indices. Sat 2005, 1, 99.

[5] Srivastava, G.; Kumar, P. Water quality index with missing parameters. Int. J. Res. Eng. Technol. 2013, 2, 609–614.

DOI: https://doi-ds.org/doilink/08.2024-48462464

[6] The Environmental and Protection Agency, "Parameters of water quality," Environ. Prot., p. 133, 2001.

[7] Kalimur Rahman, Saurav Barua, H.M. Imran, Assessment of water quality and apportionment of pollution sources of an urban lake using multivariate statistical analysis, Cleaner Engineering and Technology, Volume 5, 2021, 100309, ISSN 2666-7908

[8]Arunkumar, R & Thambusamy, Velmurugan. (2021). An Exploratory Data Analysis Process on Groundwater Quality Data. 54. 41-48

[9] Marisol Vega, Rafael Pardo, Enrique Barrado, Luis Debán, Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis, Water Research, Volume 32, Issue 12, 1998, Pages 3581-3592, ISSN 0043-1354.

[10] Boulesteix, A.L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2012.

[11]Jiang, J.; Tang, S.; Han, D.; Fu, G.; Solomatine, D.; Zheng, Y. A comprehensive review on the design and optimization of surface water quality monitoring networks. Environ. Model. Softw. 2020.

[12] C.V. Sillberg, P. Kullavanijaya, O. Chavalparit, Water quality classification by integration of attribute-realization and support vector machine for the chao phraya river, Journal of Ecological Engineering 22 (2021), 70–86.

[13] M. Yilma, Z. Kiflie, A. Windsperger, N. Gessese, Application of arti- ficial neural network in water quality index prediction: a case study in little Akaki River, Addis Ababa, Ethiopia, Modeling Earth Systems and Environment 4 (2018), 175–187.

[14] Y.R. Ding, Y.J. Cai, P.D. Sun, B. Chen, The use of combined neural networks and genetic algorithms for prediction of river water quality, Journal of Applied Research and Technology 12 (2014), 493–499.

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

[16] Zhang, J., Zhu, X., Yue, Y., & Wong, P. W. (2017). A real-time anomaly detection algorithm/or water quality data using dual time-moving windows. 2017 Seventh international conference on innovative computing technology (INTECH) (pp. 36–41). IEEE.

[17] WC Leong, A Bahadori, J Zhang, and Z Ahmad, Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM), International Journal of River Basin Management, Vol. 19, 2021, pp. 149-156.

DOI: https://doi-ds.org/doilink/08.2024-48462464