Indian Journal of Agriculture Humanity and Science
Vol. 1 Issue. 1 July – December 2025
Research Article

Assessment of Heavy Metal Contamination in Groundwater Coupled with Integrated Machine Learning-Based Arsenic Prediction

Suraj Kumar[1], Abhishek Kumar Mishra[2]& Nityanand Singh Maurya[3*]
[1, 2, 3] Department of Civil Engineering, National Institute of Technology Patna – 800005, India
*Corresponding author's *e-mail: nsmaurya@nitp.ac.in*

Received on: 28/07/2025 Revised on: 26/08/2025 Accepted: 08/09/2025

Abstract

Monitoring groundwater quality is critical for ensuring public health, environmental sustainability, and effective water resource management. Traditional monitoring methods, though reliable, are often time-consuming and lack predictive capabilities. This study applies machine learning (ML) models—Decision Tree Regressor (DTR) and Random Forest Regressor (RFR)—to predict arsenic concentration in groundwater samples collected from Bodh Gaya, Bihar, India, during the pre-monsoon season. Fifty samples were analyzed for pH and heavy metals, revealing frequent exceedance of permissible limits for Fe, Al, Mn, and As. Data preprocessing included normalization, scaling, and splitting into training and testing sets. Model performance was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Nash–Sutcliffe Efficiency (NSE). Results indicate that DTR outperformed RFR with lower error values (MSE = 0.00210, RMSE = 0.04365) and higher accuracy (NSE = 0.891, R^2 = 0.891). Feature importance analysis identified chloride as the most influential predictor, with varying contributions from other metals. The findings demonstrate the potential of ML-based approaches for real-time water quality prediction, enabling proactive interventions for sustainable groundwater management.

Keywords: Arsenic Prediction, Decision Tree Regressor, Groundwater Quality, Heavy Metal Contamination, Machine Learning, Random Forest Regressor

Introduction:

Machine learning (ML), a subset of artificial intelligence, has revolutionized data analysis by enabling systems to automatically learn from data and improve predictions without explicit programming. Its applications span diverse fields, including healthcare, finance, and environmental monitoring. ML algorithms excel in handling large datasets, identifying patterns, and making predictions, which traditional methods often struggle to achieve efficiently(Essamlali et al., 2024; Zhu et al., 2022). In environmental science, ML has emerged as a powerful tool for addressing complex challenges such as pollution detection, climate modeling, and resource management(Essamlali et al., 2024; Rajitha et al., 2024).

Environmental monitoring is critical for understanding and mitigating the impact of human activities on ecosystems. However, traditional methods of monitoring are often labour-intensive, expensive, and prone to missing subtle trends. ML offers transformative solutions by automating data analysis from sources like satellite imagery, ground sensors, and remote sensing technologies. For example, ML algorithms can predict air quality based on weather patterns and pollution sources or anticipate forest fire spread using meteorological data(Akhlaq et al., 2024; Lowe & Qin, 2022). In water quality monitoring, ML is utilized to evaluate contamination levels in surface water, groundwater, and drinking water systems. By analyzing historical data and real-time inputs, ML models can identify pollution sources and predict future contamination events(Essamlali et al., 2024; Zhu et al., 2022).ML plays a pivotal role in policy-making by providing actionable insights derived from predictive models. For instance, ML can assess the likelihood of environmental regulation violations by analyzing facility characteristics such as location and inspection history. This allows government agencies to prioritize inspections more effectively under budget constraints(Lowe & Qin, 2022; Talha et al., 2023). In water management systems, predictive models can forecast contamination risks based on rainfall patterns or industrial activities, enabling timely interventions. By leveraging these predictions, policymakers can craft regulations that are proactive rather than reactive(Zhu et al., 2022; Talha et al., 2023).

Groundwater is indispensable for life on Earth. It sustains ecosystems, supports

agriculture, drives industrial processes, and fulfils basic human needs. Despite its abundance, only a fraction of Earth's water is accessible as freshwater suitable for consumption. The importance of preserving water resources cannot be overstated as they underpin global food security, economic stability, and public health(Ashwini et al., 2019; Kalaivanan & Vellingiri, 2022).Clean water is essential for maintaining ecological balance and ensuring public health. Contaminated water poses severe risks such as the spread of diseases, disruption of aquatic ecosystems, and economic losses in industries reliant on clean water sources. Monitoring water quality helps prevent these adverse outcomes by identifying pollutants early and enabling corrective measures(Ashwini et al., 2019; Kalaivanan & Vellingiri, 2022).Water pollution arises from various sources such as industrial discharges, agricultural runoff containing pesticides and fertilizers, untreated sewage, and natural disasters like floods that carry contaminants into waterways. Polluted water affects biodiversity by disrupting aquatic habitats and threatens human health through exposure to toxins like heavy metals and pathogens(Ashwini et al., 2019; Zhu et al., 2022). Additionally, pollution exacerbates the scarcity of clean drinking water in many regions worldwide.Drinking water contamination is a pressing issue globally. Common contaminants include microbial pathogens (e.g., bacteria), chemical pollutants (e.g., fluoride, nitrates), heavy metals (e.g., arsenic, lead), and emerging pollutants like pharmaceuticals. Contaminated drinking water can cause severe health problems ranging from gastrointestinal infections to long-term illnesses like cancer(Ashwini et al., 2019; Zhu et al., 2022;Hu et al., 2023). Ensuring safe drinking water requires robust monitoring systems capable of detecting contaminants quickly and accurately.

ML has demonstrated remarkable capabilities in predicting groundwater contamination events and identifying probable causes. By analysing complex datasets—such as weather conditions, industrial activity logs, or historical contamination records—ML models can forecast pollution trends with high accuracy(Essamlali et al., 2024; Zhu et al., 2022; Vora et al., 2025; Xu et al., 2022). For example:Surface Water: ML algorithms predict changes in surface water quality based on urban wastewater discharge patterns.Groundwater: Groundwater safety assessments use ML to identify pollution sources like agricultural runoff.Drinking

Water: Advanced models forecast microbial contamination risks during heavy rainfall events.These predictions enable stakeholders to implement preventive measures before contamination becomes widespread.

Machine learning represents a paradigm shift in environmental monitoring and management. Its ability to analyze vast datasets efficiently makes it invaluable for predicting environmental changes and informing policy decisions. In the context of water quality management, ML provides critical insights into contamination risks across various water types—surface water, groundwater, drinking water—and their probable causes. As global challenges like climate change intensify pressures on freshwater resources, adopting ML-driven solutions will be essential for safeguarding this vital resource while ensuring sustainable development.

2. Material and methods

2.1. Study area

The study was conducted in Bodh Gaya, Gaya district of Bihar state, India. Gaya is located at 24. 47' N latitude and 84. 50' E longitude, with an average elevation of 113 meters above sea level. Arwal, Jehanabad, Nalanda, Nawada, and Aurangabad districts delineate the northern, eastern, and western boundaries of the region, while Jharkhand lies to the south. The area's drainage system is governed by four principal rivers: the Morhar, Phalgu, Paimar, and Dhadhar. Originating from the southern plateau of Jharkhand, these rivers flow predominantly in a north and north-easterly direction through the study region.The entire study area experiences a continental monsoon climate with challenging environmental conditions. Summers are marked by intense westerly winds, with temperatures soaring up to 46 °C, while winter temperatures can drop as low as 4 °C. The monsoon season typically starts in late June and continues into early July, with annual rainfall ranging between 568.5 and 1,109 mm (CGWB 2022).

2.2.Sample collection and laboratory analysis

Total 50 (Fig. 1) groundwater samples were collected from the selected area of Bodh Gaya, Gaya district from hand-pump as well as borewells during the pre-monsoon season, which ranged in depth from 12 to 35 meters. Each sample was obtained using clean, dry polypropylene bottles and kept in a cool, dry container until laboratory analysis (APHA 2017). Before sampling, the pumps and wells were purged for 5–10 minutes to remove stagnant water and ensure samples

were representative. The bottles were rinsed 2–3 times with the collected groundwater prior to filling.Each 125 mL sample was filtered through a 0.4 µm Millipore membrane. The filtered samples were acidified with HNO_3 to reduce the pH to 2. pH measurements were taken with a portable meter (PCS Testr 35). Heavy metal concentrations were determined using inductively coupled plasma optical emission spectrometry (ICP-OES, Agilent 5110 VDV), with the equipment operated according to the manufacturer's guidelines for maximum accuracy. To validate the analytical results, each sample was measured in triplicate. All reagents were of analytical grade, and blanks and standard solutions were regularly analyzed after every tenth sample to ensure precision and reliability.

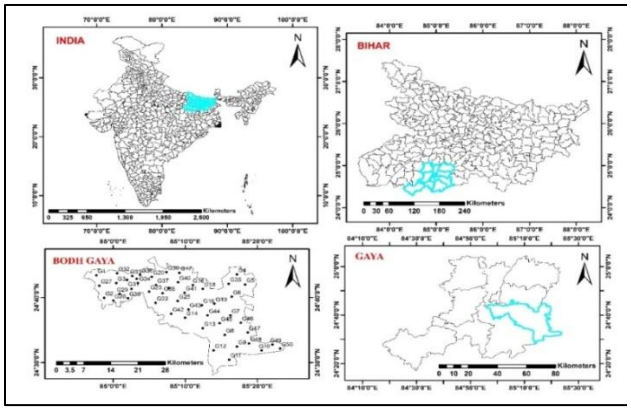


Fig 2: Study area map (Bodh Gaya) and location of sampling

2.3 Data processing

The process of applying machine learning prediction begins with the Python language, where the code is written and accessible through the sklearn library. Next, the libraries and data are imported, and the data is pre-processed before any machine learning model is applied. This includes standardization for classification and normalization/scaling for prediction; the application of either or both of these models depends on the requirements and the type of data, as well as data splitting for training and testing.In order to improve the machine learning model's applicability, normalization/scaling is used to transform features on a similar scale (0-1) for uniformity in the dataset. It can be significantly impacted by outliers and is helpful for data whose distribution is unknown. Normalization/scaling is typically used in prediction modeling (Eqn. 1). On the other hand, normalization is the process of transforming features using the dataset's mean and standard deviation, which are not constrained by a range and are significantly less impacted by outliers. Standardization is employed in classification modeling and is helpful when the data distribution type is gaussian or normal (Eqn. 2) (Ibrahim et al., 2022). Based on the dataset type and the requirement, any one of the normalizations

or standardizations is applied, and after that, data splitting is done.

$$X_{Scale} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

$$X_{Stand.} = \frac{X - mean}{Std.Dev.} \quad (2)$$

The process of data splitting is used to train and test the machine learning model; essentially, the data is split into two halves (either in an 80/20 or 70/30 ratio). The model learns the data structure, behaviour, interlinking, parameter relations, and patterns from the training data. Following the model's application, the data is predicted. The testing data is then compared to the anticipated data, and a number of evaluations are conducted to determine the model's accuracy, behaviour, inaccuracy, and suitability for a given data type (Singh et al., 2022).

After data processing, a specific model is selected and applied, and various model assessment metrics are calculated and results are interpreted based on prediction data**.**

2.4 Metrics for assessment

Various model assessment metrics such as Mean square error (MSE), Root mean square error (RMSE), Nash - Sutcliffe efficiency (NSE), and Coefficient of determination (R^2) were employed, and can be expressed mathematically by Eqn. 3-6 (Ibrahim et al., 2022).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \quad (4)$$

$$R^2 = \frac{((\sum_{i=1}^{n}(o_i - o_m)p_i - p_m))^2}{\sum_{i=1}^{n}(o_i - o_m)^2 * \sum_{i=1}^{n}(p_i - p_m)^2} \quad (5)$$

$$NSE = \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(o_i - o_m)^2} \quad (6)$$

Where, n is the total number of test samples, o_i represents the observed value, p_i represents the predicted value, $o_m$represents the mean of the observed values and p_m is the mean of the predicted values

2.5 Model development and application

. **Random Forest Regressor (RFR)**

The Random Forest Regressor (RFR) is an ensemble technique that builds a large number of decision trees using random sampling with replacement in order to produce repeating predictions of the target variable. According to Ibrahim et al. (2022), the method, which is seen in Fig. 2, uses the performance of several decision tree algorithms to forecast arsenic concentration. A training subset chosen at random with replacement is used by each decision tree,

and this subset is repeated as many times as there are trees in the ensemble. A final forecast is then generated by combining the results of various decision trees (Zhou et al., 2024). Using bootstrap sampling, a random subsample of the training dataset is used to generate each tree; the samples that are excluded from this subsample are known as "Out of Bag" (OOB) samples (Jin et al., 2020). The RFRM model is internally cross-validated using these OOB samples (Chakraborty et al., 2020). Hyperparameter tweaking was used to improve the prediction of arsenic concentration levels after the initial RFRM model was developed.

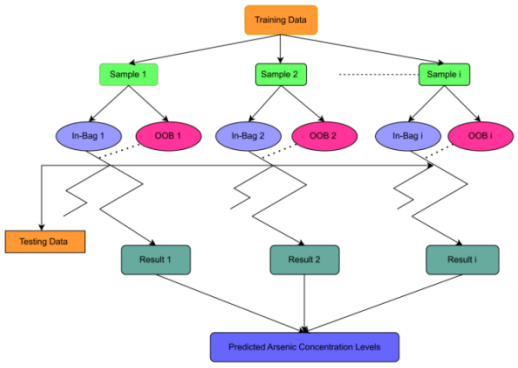


Fig. 2: Flow chart of Random forest regressor algorithm

Decision Tree Regressor (DTR)

Regression and classification problems are addressed using non-parametric machine learning methods known as decision trees (DTs). Unlike black-box algorithms, DTs feature an easy-to-understand decision-making process and are highly intuitive (Ibrahim et al., 2022). Before rendering decisions based on a range of input variables organized into layers of decision branches, the algorithm begins at a root node and proceeds through internal and terminal nodes. Following the initial split of the data into two subsets, the decision tree (DT) uses the same logic to split each subsequent subset recursively.This procedure keeps on until either the maximum depth specified is achieved or no further splits that minimize the loss function can be identified (Singh et al., 2022). Decision trees (DTs) are widely used classifiers and regressors for developing binary classification and regression, respectively, because of their simplicity and interpretability. Compared to other algorithms, DTs handle numerical and categorical data efficiently and with fewer assumptions.

3 Results and discussions

3.1 Analysis of heavy metals

Assessing the presence of heavy metals in groundwater and their absorption via many exposure routes may increase the risks to human health. In accordance with BIS 10500:2012, Table 1 lists the permitted limits for the pH and heavy metal concentrations as well as the statistical data

(maximum, minimum, standard deviation, and third quartile). According to the third quartile, the research reveals that the heavy metals' abundance is as follows: Fe> Zn> Mn> Al> Cu>As>Cr> Ni> Pb>Cd. Both natural and man-made causes, such as increased concentrations of $CO_3^{2-}$and SO_4^{2-} minerals and other minerals, may contribute to the presence of heavy metals in groundwater. Anthropogenic sources include mining, agriculture, and industrial and household wastewater (Liu et al. 2019; Kozisek 2020).

A detailed assessment of groundwater samples has revealed the presence of several heavy metals, many of which exceed their respective permissible limits, indicating significant contamination in the studied region. The analysis found that aluminum (Al) concentrations in the samples ranged from 12.01 to 445.26 µg/L. Alarmingly, 56% of these samples contained aluminum levels above the established safety threshold. Elevated aluminum can be toxic to both humans and aquatic life, posing health risks when present in drinking water supplies.

Arsenic (As), another toxic element, was detected in a range spanning from non-detectable (ND) amounts up to 15.48 µg/L. Despite the relatively narrow range, 10% of the samples still surpassed the maximum permissible limit for arsenic, a contaminant associated with a variety of chronic health issues—including skin lesions, cancer, and cardiovascular diseases—when consumed over extended periods.Cadmium (Cd) analysis showed concentrations varying from ND to 1.21 µg/L. Fortunately, in this survey, none of the assessed samples exceeded the recognized safe limit for cadmium. It is important to note, however, that cadmium's presence—even at low concentrations—warrants attention, as it is highly toxic and accumulates over time in living organisms. The primary sources of environmental cadmium include the combustion of fossil fuels, municipal waste incineration, the improper disposal of nickel–cadmium (Ni–Cd) batteries, and various industrial processes, as highlighted by the Agency for Toxic Substances and Disease Registry (ATSDR 2017).Chromium (Cr) was measured in concentrations ranging from 0.94 to 46.54 µg/L. Like many heavy metals, chromium contamination can arise from industrial activities, though the paragraph does not indicate how many samples exceeded permissible limits for chromium.Iron (Fe), recognized as an essential micronutrient for numerous

organisms, was found in concentrations stretching from 56 to a remarkable 12,350 µg/L—revealing extreme variation across the sampling sites. Over half (58%) of the collected samples contained iron levels above the acceptable threshold. While iron is required for vital metabolic processes, its excess—mostly stemming from the weathering of igneous rocks, ferromanganate soils, or the dissolution of minerals such as iron oxides, magnetite, sulfides, and iron clays (Ranjan et al. 2012)—can cause problems like taste, staining, and infrastructural damage in water distribution systems.

Manganese (Mn), also crucial for human health as a micronutrient, exhibited concentrations between 1.42 and 1,430 µg/L in the samples. Around one-fifth (20%) of the tested groundwater samples exceeded safe manganese levels. Sources may be both natural (geological formations) as well as anthropogenic (industrial or agricultural activities), and high levels of manganese can result in neurological health effects if ingested over time.Nickel (Ni) is a heavy metal known for its carcinogenic potential, implicated in causing serious lung and kidney diseases (Multhaup 2005). In this study, nickel concentrations ranged from non-detectable up to 33.96 µg/L, with approximately 4% of samples exceeding the permissible limit. Industrial activities, including operations at power plants, glass and ceramics manufacturing, battery production, dye and colorant industries, and urban sewage and sludge, are recognized sources of groundwater nickel contamination (Govil et al. 2008).Lead (Pb), primarily an anthropogenic pollutant due to its minimal natural release relative to extensive human-associated discharges (Buragohain et al. 2010), was found at concentrations ranging up to 22.99 µg/L in the analyzed region. About 8% of the samples contained lead levels above recommended standards. Major local sources of lead include the historical usage of leaded gasoline, emissions from paint industries, the application of synthetic pesticides, fossil fuel combustion, and poor management of battery waste (Jumbe & Natural 2009).Zinc (Zn), an essential trace element, showed a wide range of concentrations from 12.18 up to a substantial 6,280.55 µg/L. However, only 2% of all samples tested contained zinc above the acceptable limit. Although typically less toxic compared to other heavy metals, excess zinc in drinking water may impart an undesirable taste and lead to health issues if consumed at very high

levels.Collectively, these findings point to considerable variability and frequent exceedance of permissible levels for several key heavy metals in the groundwater of the studied region, underlining an urgent need for ongoing monitoring, mitigation, and public health interventions.

2.3.Performance metrics for arsenic prediction

The performance comparisonfor arsenic prediction (Table 2) indicates that the Decision Tree Regressor (DTR) outperforms the Random Forest Regressor (RFR) for the testing dataset. DTR shows lower MSE (0.00210) and RMSE (0.04365), along with higher NSE and R^2 values (both 0.891), suggesting better prediction accuracy and model efficiency. In contrast, RFR yields comparatively higher error values and lower correlation with actual observations. These results highlight DTR as the more suitable model for this dataset.

The feature importance plots (Fig. 3) for DTR and RFR highlight Chloride as the most significant predictor of the target variable in both models. The most significant feature in the RFR plot (green bars) is PH, which is closely followed by Mn and Cr. The ensemble-based RFR model identified these three traits to be consistently predictive across its constituent decision trees, as evidenced by their significantly higher normalized significance values than the others. Cd, Fe, and Pb have very little contribution, indicating little or no predictive potential in the context of this model, but EC, Cu, and Al are other fairly significant features. On the other hand, a distinct ranking is revealed by the DTR plot (red bars). The most significant feature is Cr, which is closely followed by Mn, Fe, and PH. It's interesting to note that whereas Fe is seen as highly significant in DTR, it is almost meaningless in RFR. The reason for this disparity is that decision trees may be overly focused on a small number of variables that produce significant initial splits and may be sensitive to even little changes in the data. Furthermore, in comparison to RFR, DTR gives a comparatively greater weight to a wider variety of characteristics, such as Zn, Ni, and Cd.

These variations draw attention to a crucial aspect of RFR: it reduces the variance and bias brought forth by individual trees by averaging out feature importance over several trees. Because DTR is a single-tree model, it may overfit and highlight features that best split the training data but may not generalize well, whereas RFR offers a more

reliable and broadly applicable interpretation of feature value.

3. Conclusions

This study demonstrates that machine learning offers an efficient, accurate, and interpretable method for groundwater quality prediction. The Decision Tree Regressor (DTR) outperformed the Random Forest Regressor (RFR) in predicting arsenic concentrations in the Bodh Gaya region, achieving higher accuracy and lower error values. Chloride emerged as the most significant predictor, alongside other influential variables such as manganese, chromium, and pH. The groundwater analysis revealed elevated concentrations of several heavy metals, with Fe, Al, Mn, and As frequently exceeding permissible limits, posing potential health and environmental risks. By integrating ML models into groundwater monitoring frameworks, water resource managers can make proactive, data-driven decisions, enabling timely mitigation strategies to safeguard public health and ensure sustainable water management.

ACKNOWLEDGEMENT

Authors are thankful to Dr Sumant Kumar and Director, NIH Roorkee for providing laboratory analysis facilities.

FUNDING

No funding was obtained/provided from anywhere for this study.

AUTHORS CONTRIBUTION

S.K. and A.K.M contributed to the literature review, conceptualization, sample collection, laboratory analysis, results analysis, and interpretation,writing the original manuscript. N.S.M. contributed to conceptualization, methodology development, and resultsinterpretation, supervision/guidance, and reviewed the original manuscript

STATEMENTS & DECLARATIONS

All authors have read, understood, and have complied as applicable with the statement on ‘Ethical responsibilities of Authors’as found in the Instructions for Authors.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict

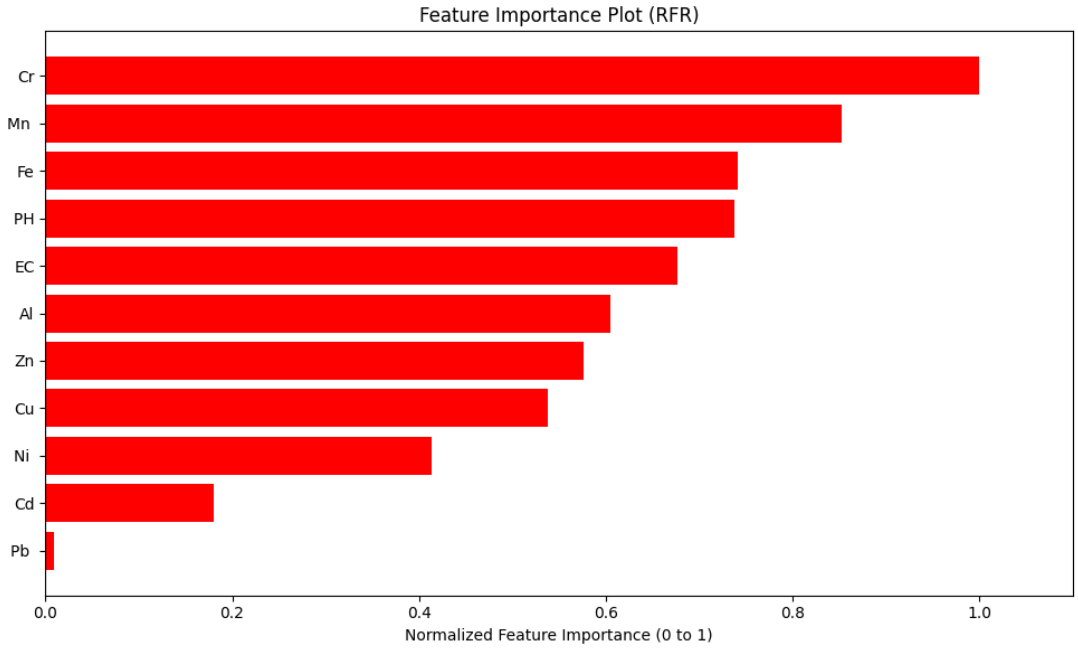
Table 1. Statistical summary of heavy metal concentration compared with BIS (10500: 2012)

Elements	Range	Mean ± Std. deviation	3rd Quartile	BIS Standard (10500 2012)		% of the sample exceeding the Acceptable limit (BIS10500)
				Acceptable limit	Permissible limit	
pH	6.82-8.25	7.34±0.23	7.49	6.5-8.5	-	-
EC	223-1357	609.04±276.67	828.25	-	-	-
Al	12.01-445.26	49.92 ± 66.71	46.375	30	200	56
As	0 - 15.48	2.98 ± 4.38	6.5425	10	50	10
Cd	0 - 1.21	0.09 ± 0.26	0	3	NR	0
Cr	0.94 - 46.54	3.94 ± 6.51	3.6525	50	NR	0
Cu	0 - 151.59	13.65 ± 30.42	6.895	50	1500	8
Fe	56 - 12350	1254.8 ± 1946.32	1560	300	NR	58
Mn	1.42 - 1430.69	88.64 ± 211.65	60.0625	100	300	20
Ni	0 - 33.96	2.94 ± 6.42	2.5225	20	NR	4
Pb	0 - 22.99	1.71 ± 5.32	0	10	NR	8
Zn	12.18 - 6280.55	556.83 ± 1183.67	302.6425	5000	15000	2

***Heavy metals are measured in µg/l., NR – No relaxation**

Table 2: Performance of ML models

Models	**Performance criteria (Accuracy for testing sets)**			
	MSE	RMSE	NSE	R^2
DTR	0.00210	0.04365	0.891	0.891
RFR	0.00671	0.08123	0.856	0.856



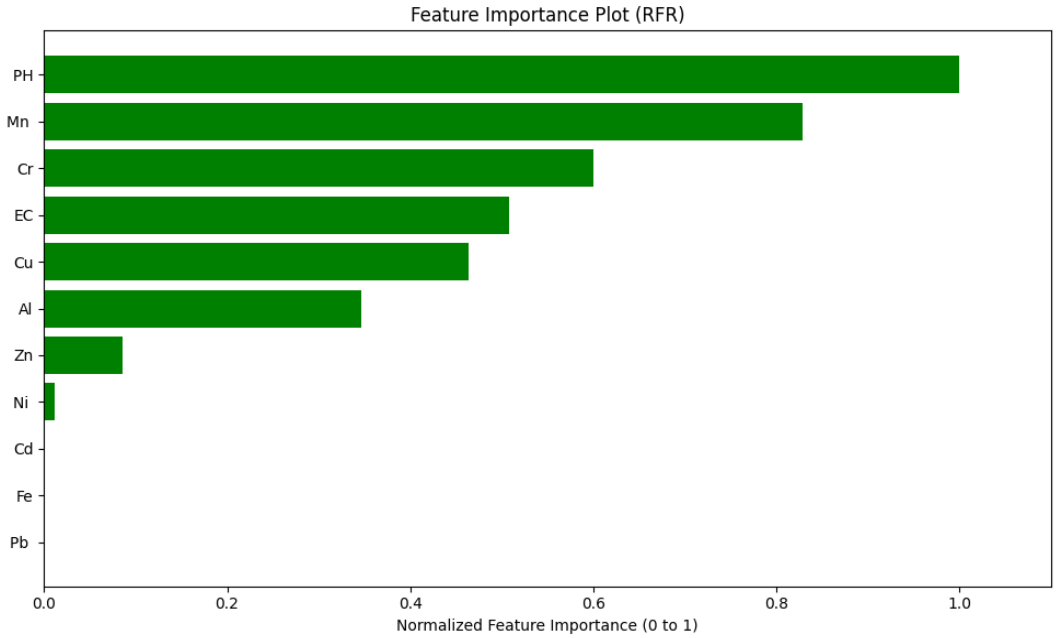


Fig 3: Feature importance plots for (a) RFR, (b) DTR

References

Akhlaq, M., Ellahi, A., Niaz, R., Khan, M., Sammen, S. S., & Scholz, M. (2024). Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction. *Tellus A: Dynamic Meteorology and Oceanography*, 76(1), 177–192. https://doi.org/10.16993/tellusa.4069

Ashwini, C., Singh, U. P., Pawar, E., & Shristi. (2019). Water quality monitoring using machine learning and IoT. *International Journal of Scientific and Technology Research*, 8(10), 1046–1048.

ATSDR. (2017). *Appendix a. Atsdr Minimal Risk Levels and Worksheets*. 1–32.

Buragohain, M., Bhuyan, B. & Sarma, H. P. (2010) Seasonal variations of lead, arsenic, cadmium and aluminium contamination ofgroundwater in Dhemaji district, Assam, India, Springer, 170 (1–4), 345–351. https://doi.org/10.1007/s10661-009-1237-6.

Chakraborty, M., Sarkar, S., Mukherjee, A., Shamsuddhuha, M., Ahmed, M. K., Bhattacharya, A., & Mitra, A. (2020). Modeling Regional-Scale Groundwater Arsenic Hazard in the Transboundary Ganges River Delta, India and Bangladesh: Infusing Physically-Based Model with Machine Learning. *Science of the Total Environment*, 748, 141107. https://doi.org/10.1016/j.scitotenv.2020.141107

Essamlali, I., Nhaila, H., & El Khaili, M. (2024). Advances in machine learning and IoT for water quality monitoring: A comprehensive review. *Heliyon*, 10(6), e27920. https://doi.org/10.1016/j.heliyon.2024.e27920

Govil, P. K., Sorlie, J. E., Murthy, N. N., Sujatha, D., Reddy, G. L. N., Rudolph-Lund, K., Krishna, A. K. & Rama Mohan, K. (2008) Soilcontamination of heavy metals in the Katedan Industrial Development Area, Hyderabad, India, Environmental Monitoring andAssessment, 140 (1–3), 313–323. https://doi.org/10.1007/s10661-007-9869-x.

Hu, X. C., Dai, M., Sun, J. M., & Sunderland, E. M. (2023). The Utility of Machine Learning Models for Predicting Chemical Contaminants in Drinking Water: Promise, Challenges, and Opportunities. *Current Environmental Health Reports*, 10(1), 45–60. https://doi.org/10.1007/s40572-022-00389-x

Ibrahim, B., Ewusi, A., Ahenkorah, I., &Ziggah, Y. Y. (2022). Modelling of arsenic concentration in multiple water sources: A comparison of different machine learning methods. *Groundwater for Sustainable Development*, https://doi.org/10.1016/j.gsd.2022.100745

Jin, Z. (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *12343 LNCS*, 503–515.

Jumbe, A. & Natural, N. (2009) Impact assessment of heavy metals pollution of Vartur lake, Bangalore, Journal of Applied and Natural Science,1 (1), 53–61.Available at: http://journals.ansfoundation.org/index.php/jans/article/view/35.

Kalaivanan, K., &Vellingiri, J. (2022). Survival Study on Different Water Quality Prediction Methods Using Machine Learning. *Nature Environment and Pollution Technology*, 21(3), 1259–1267. https://doi.org/10.46488/NEPT.2022.v21i03.032

Kozisek, F. (2020) Regulations for calcium, magnesium or hardness in drinking water in the European Union member states, RegulatoryToxicology and Pharmacology, 112, 104589.https://doi.org/10.1016/J.YRTPH.2020.104589.

Liu, X., Ma, R., Wang, X., Ma, Y., Yang, Y., Zhuang, L., Zhang, S., Jehan, R., Chen, J. & Wang, X. (2019) Graphene oxide-based materials forefficient removal of heavy metal ions from aqueous solution: a review, Environmental Pollution, 252, 62–73. https://doi.org/10.1016/J.ENVPOL.2019.05.050.

Lowe, M., & Qin, R. (2022). A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring. *Water*, 14(1384), 1–28.

Rajitha, A., Aravinda, K., Nagpal, A., Kalra, R., Maan, P., Kumar, A., & Abdul-Zahra, D. S. (2024). Machine Learning and AI-Driven Water Quality Monitoring and Treatment. *E3S Web of Conferences*, 505, 1–11. https://doi.org/10.1051/e3sconf/202450503012

Ranjan, R. K., Ramanathan, A. L., Parthasarathy, P. & Kumar, A. (2012) Hydrochemical characteristics of groundwater in the plains ofPhalgu River in Gaya, Bihar, India, Arabian Journal of Geosciences, 6 (9), 3257–3267. https://doi.org/10.1007/s12517-012-0599-1.

Multhaup, G. (2005) Environmental copper and manganese in the pathophysiology of neurologic diseases (Alzheimer's disease andmanganism) article in acta hydrochimica et

hydrobiologica · April 2005, Wiley Online Library, 33 (1), 72–78. https://doi.org/10.1002/aheh.200400556.

Singh, S. K., Taylor, R. W., & Pradhan, B. (2022). Predicting sustainable arsenic mitigation using machine learning techniques. *Ecotoxicology and Environmental Safety*, https://doi.org/10.1016/j.ecoenv.2022.113271

Talha, M., Nagpal, N., & Srivastava, M. (2023). Applying Machine Learning for Ensuring Sustainable Management of Water (SDG6). *CEUR Workshop Proceedings*, 3619, 42–56.

Vora, K. B., Mashru, D. V., Doshi, S. M., & Bhalodiya, V. V. (2025). The Role of Machine Learning in Water Quality Assessment: Current Applications and Future Scope. 1–6. https://doi.org/10.55041/IJSREM41578

Xu, X., Lai, T., Jahan, S., Farid, F., & Bello, A. (2022). A Machine Learning Predictive Model to Detect Water Quality and Pollution. *Future Internet*, 14(11), 1–14. https://doi.org/10.3390/fi14110324

Zhou, X., Su, C., Xiajun, X., Ge, W., Xiao, Z., Yang, L., & Pan, H. (2024). Employing Machine Learning to Predict the Occurrence and Spatial Variability of High Fluoride Groundwater in Intensively Irrigated Areas. *Applied Geochemistry*, 167(April), 106000. https://doi.org/10.1016/j.apgeochem.2024.106000

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment and Health*, 1(2), 107–116. https://doi.org/10.1016/j.eehl.2022.06.001

.**How to cite this article:** Kumar, S., Mishra, A. kumar ., & Maurya, N. S. (2025). Assessment of Heavy Metal Contamination in Groundwater Coupled with Integrated Machine Learning-Based Arsenic Prediction. Indian Journal of Agriculture Humanity and Science, 1(1), 32–44. https://doi.org/10.5281/zenodo.17202004