PREDICTION OF HEAVY METAL CONTAMINATION USING MACHINE LEARNING ALGORITHMS IN THE MIDDLE GANGA BASIN

A thesis submitted in partial fulfillment of the requirements for the award of the degree of

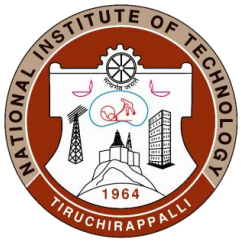
B. Tech.

in

CIVIL ENGINEERING

By

HARSH KUMAR BHATT (103119042)
HIRDYANSH RASTOGI (103119044)
KAPIL KUMAR BAIRWA (103119050)
RAJNESH MEENA H (103119086)



DEPARTMENT OF CIVIL ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
TIRUCHIRAPPALLI-620 015

MAY 2023

BONAFIDE CERTIFICATE

This is to certify that the project titled **PREDICTION OF HEAVY METAL CONTAMINATION USING MACHINE LEARNING ALGORITHMS IN THE MIDDLE GANGA BASIN** is a bonafide record of the work done by

HARSH KUMAR BHATT (103119042)

HIRDYANSH RASTOGI (103119044)

KAPIL KUMAR BAIRWA (103119050)

RAJNESH MEENA H (103119086)

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **CIVIL ENGINEERING** of the **NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI**, during the year 2022-23.

Dr. R. GANDHIMATHI
Project Guide

Dr. S.T. RAMESH
Head of the Department

Project Viva-voce held on _____

Internal Examiner

External Examiner

ABSTRACT

Heavy Metal (HM) contamination is a major environmental issue that poses a serious threat to human health and ecosystems. The Middle Ganga Basin in India has been particularly affected by heavy metal pollution due to the discharge of untreated industrial waste into water sources and the major consequence we can observe is the degradation of the water quality. This study intends to analyse the effect of Heavy Metal on Water Quality and predict their concentration using Machine learning algorithms in Middle Ganga Basin encompassing Uttar Pradesh (UP) state, India. Evaluation of the HM concentration in the rivers is quite difficult and requires more time and effort. The study analyzed six heavy metals, including arsenic, cadmium, chromium, nickel, lead, and zinc, to identify the most significant predictors of heavy metal contamination. The study examined the spatial distribution of water quality index (WQI) and heavy metal concentrations using geographic information system (GIS) mapping, which revealed the relationship between hotspots of contamination and WQI in specific areas. Water quality data from 44 water quality monitoring stations in the basin, including physicochemical parameters and heavy metal concentrations, to train and evaluate the performance of four machine learning algorithms: Linear Regression, Random Forest, Artificial Neural Network, and Gradient Boosting Regression. The XGBoost model was found to be the best-performing algorithm, with an accuracy of over 98%, making it an effective tool for predicting heavy metal contamination in the region. The developed model can help policymakers and researchers develop targeted intervention strategies to mitigate the impact of heavy metal pollution in the Middle Ganga Basin.

Keywords: Heavy Metal Contamination, Water Quality, XGBoost Model, ANN

ACKNOWLEDGEMENT

We would like to thank the following people for their support and guidance without whom the completion of this project in fruition would not be possible.

We feel it is a great privilege in expressing my thanks to my guide **Dr. R. Gandhimathi**, Professor, Department of Civil Engineering for her valuable support and guidance in making this project a successful one.

We wish to express my gratitude and thanks to **Dr. S. T. Ramesh,** Professor and Head, Department of Civil Engineering for his full-fledged support and motivation.

We deem it a dip in delight and ecstasy to thank **Dr. G. Aghila,** Director, National Institute of Technology, Tiruchirappalli, for having provided us with sufficient facilities to carry out this project.

Our internal reviewers, **Dr. Nisha Radhakrishnan**, **Dr. Sevugan Rajkannu J**, **Dr. Chandaluri Vinaykumar**, **Dr. Laiju A. R**, and **Dr. Makendran C** for their insight and advice provided during the review sessions.

We would also like to thank our parents and friends for their constant support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Water pollution is one of the major environmental concerns worldwide and poses a significant threat to human health, aquatic life, and ecosystems. Chemical pollutants are one of the leading causes of water pollution, and heavy metals are among the most toxic and persistent of these pollutants (Magdaleno et al. 2014). Heavy metal pollution in water can be caused by natural sources such as weathering of rocks or anthropogenic sources such as mining, industrial discharge, wastewater treatment plants, and agricultural runoff (Concas et al., 2006; Basak and Alagha, 2010; Sylaios et al., 2012; Ghosh and Maiti, 2018).

Heavy metals such as lead, arsenic and cadmium are as being of major public concern regarding their toxicity (WHO, 2017). These metals can accumulate in the body over time and have toxic effects on various organs and systems. Heavy metal pollution in water has been linked to a wide range of health problems, including cancer, neurological disorders, and developmental abnormalities (Dokmeci 2017; Aricak et al. 2019; Aricak et al. 2020; Sevik et al. 2020). Knowing heavy metals concentrations and behaviors is thus essential for environmental management, especially when it comes to protecting safe drinking water sources Alizamir and Sobhanardakani 2017; Lu et al. 2019).

Numerous studies have examined the levels and hazards of dissolved heavy metals in lakes and rivers (Muller et al., 2008; Jiang et al., 2012; Deng et al., 2018; Liang et al., 2018; Rajeshkumar et al., 2018). The extent of water pollution caused by heavy metals depends on the location, pollution source, and the types of heavy metals present. It has been discovered that particulate heavy metals can also bioaccumulate (Bourgeault et al., 2011; Sevik et al. 2019; Cetin 2019). Additionally, research has identified various physicochemical and environmental factors, such as pH, organic matter, suspended matter, and dissolved oxygen level, that can influence heavy metal behavior (Cardwell et al., 2013; La Colla et al., 2015; Wang et al., 2016; Yang et al., 2016). Consequently, continual monitoring of heavy metal levels and other environmental parameters is crucial for a comprehensive understanding of heavy metal behavior and its associated

risks in aquatic environments.

Measurement of heavy metals in water is a challenging task due to the complex nature of water samples, the low concentration of heavy metals, and the interference from other compounds. Developed analytical techniques can be used to measure heavy metal concentrations in water, including aqua regia digestion, atomic absorption spectrometry (AAS) and inductively coupled plasma mass spectrometry (ICP-MS). However, these techniques are expensive and require specialized equipment and expertise, making them difficult to implement in areas with limited resources. Moreover, these methods do not provide real-time monitoring of heavy metal contamination, making it difficult to respond quickly to contamination events. Heavy metal concentrations and other environmental indices should be constantly monitored to comprehensively understand the behaviour and associated risks of heavy metals in aquatic environments; however, it should be reiterated that regular and integrated monitoring is time consuming and costly, and integrated monitoring systems that consider both heavy metals and other relevant environmental indices are often lacking (Lu et al. 2019).

Due to their relatively simple concepts and ease of implementation, linear models have been the focal point of many investigations and the critical tool for time series modeling over the past few decades. However, in most real problems, especially in the field of river water quality modeling, we encounter non-linear patterns that need non-linear models to deal with that. Therefore, to overcome the limitation of such linear models, in the past two decades, several non-linear models have been proposed in the literature. Among all the nonlinear techniques Machine learning techniques have gained the most popularity. Infact, in recent years, machine learning (ML) techniques have been successfully applied in various fields, including medicine, finance, and engineering, and have the potential to revolutionize environmental monitoring and management (Rozos, 2019; Kim et al., 2019). Many researchers are seeing ML techniques as a powerful tool for analyzing and predicting complex environmental data. ML algorithms can analyze large datasets and identify patterns and trends that are difficult to detect using traditional statistical methods. Globally, several researchers have applied ML techniques such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Artificial Neural Network (ANN) in various water research studies. This indicate a necessity to investigate the novel data-intelligence models which are extremely

significant due to their superior resilience and dependability in anticipating HM, while also minimizing costs and time consumption.

Because of the danger they pose, heavy metals should be investigated in drinking and utility waters. The Middle Ganga Basin, the subject of this study, is located in the northern part of India and is home to over 400 million people. The Middle Ganga Basin in India encompasses major economic centres, experiencing rapid urbanization in recent years. These developmental activities have severely impacted the basin by heavy metal pollution. The basin is a major agricultural and industrial region, and heavy metal pollution in water is a significant environmental concern. Several studies have reported high levels of heavy metal pollution in water sources in the Middle Ganga Basin, with lead and chromium being the most commonly reported heavy metals (Paul. D, 2017). To put it another way, the socio-economic viability of a state is particularly vulnerable to water quality altercations, and this sensitivity is heightened considering the fact that much agricultural operations take place on river-irrigated land. There is a need for a comprehensive investigation because only a portion of basin has been examined thus far (Charan et al., 2014; Shukla et al., 2020).

1.2 PROJECT MOTIVATION

- Heavy metal contamination is a serious environmental issue that can have adverse health effects on humans and other organisms.

- The Middle Ganga Basin is a heavily populated area with numerous industrial and agricultural activities, making it particularly susceptible to heavy metal contamination.

- With increase in heavy metal pollution, the analysis of heavy metal contamination using machine learning algorithms can also aid in the incorporation of heavy metal concentration in the calculation of water quality indices (WQI), which are commonly used to evaluate the overall quality of water resources

- Traditional methods of monitoring heavy metal contamination can be time-consuming and expensive. Machine learning algorithms offer a more efficient and accurate alternative.

- The use of machine learning algorithms can help identify potential sources of contamination and develop effective strategies for managing and mitigating heavy metal pollution in the Middle Ganga Basin.

1.3 OBJECTIVES OF THE STUDY

1. To estimate the Water Quality Index and compare hotspots of contamination with WQI,
2. To predict the concentration of heavy metals in the middle Ganga Basin by using machine learning models,
3. To identify the optimum approaches and input parameters for each heavy metal prediction with the best performance, and
4. To analyse and compare the performance of conventional and hybrid models based on the effect of input parameter selection.

1.4 THESIS ORGANIZATION

This thesis is divided into four further chapters as follows.

(a) Chapter Two: This chapter focuses on an extensive literature review focusing on thewater quality estimation and WQI, Heavy Metals contamination and Machine Learning Models.

(b) Chapter Three: The primary focus of this chapter is to delve into the methodology of this study. It explores the methodology flowchart, study area, data collection and pre-processing, WQI calculations, machine learning models developed and prediction metrices used to accurately capture the performance of the models.

(c) Chapter Four: In this chapter, the results obtained from the machine learning models deployed in the previous chapters are presented and discussed. It highlights the relationship ofHM contamination and WQI values in the study area, providinginsights into overall performance of various models.

(d) Chapter Five: This concluding chapter summarizes the key findings and conclusionsdrawn from the research work. It also discusses the practical implications of the study and offers recommendations for future research directions.

CHAPTER 2

LITERATURE REVIEW

2.1 WATER QUALITY ESTIMATION AND WQI

Surface water pollution is posing a serious challenge for water quality management (Zhang et al., 2021a). Assessing water quality is essential for water resource management (Renouf et al., 2017; Salerno et al., 2018). Water quality estimation is an important aspect of environmental monitoring. Various characteristics of surface water need to be considered during the formulation of water resource management plans. The contamination of water sources poses a serious threat to the environment and human well-being, as reported in recent research studies (Le Moal et al., 2019). As a result, several indices have been established to measure the quality of surface water, such as Water Quality Indices (WQIs), Trophic Status Indices (TSIs), and Heavy Metal Indices (HMIs), which are based on Water Quality Parameters (WQPs). These indices are designed to effectively evaluate the quality of surface water and provide guidelines for water resource management.

The initial Water Quality Index (WQI) was developed by integrating the physical and chemical characteristics of water bodies, as stated in the works of Horton (1965) and Hurley et al. (2012). The WQI is a valuable tool that presents a more accurate representation of water quality variations in particular regions, and it is an effective means of depicting water quality, as affirmed by Rangerti et al. (2015) and Tyagi et al. (2013). Nonetheless, there is no universal WQI that can be utilized to assess surface water quality, even though several modifications have been proposed to create different WQIs that are based on the specific conditions of certain areas, as noted by Sutadian et al. (2016) and Tyagi et al. (2013).

The National Sanitation Foundation (NFS) initially presented the WQI, while the Canadian Council of Ministers of the Environment (CCME) suggested another form of the WQI. According to CCME (2001) and Noori et al. (2019), various other WQIs have been established, which have been based on NFS and CCME WQIs. Some WQIs have been revised or enhanced by researchers such as Bhateria and Jain (2016), Gao et al. (2020), Khan and Jhariya (2017), and Sutadian et al. (2016).

The United States Environmental Protection Agency (USEPA) introduced the most commonly used approach for analyzing the risk of exposure to heavy metals, as reported in previous studies (Alves et al., 2014; USEPA, 1989). Evaluating the potential threat of heavy metals to human health involves the assessment of oral intake and absorption through the skin by utilizing parameters such as average daily dose (ADD), hazard quotient (HQ), hazard index (HI), and carcinogenic risk (CR) estimation, as noted in recent research (Alver, 2019; Ustaoğlu et al., 2021). Furthermore, the heavy metal indices (HMIs) have been derived from water quality indices (WQIs) and have been implemented to categorize water quality based on pollution indicators and geographical location (Gad et al., 2021). Depending on the specific conditions prevalent in an area, various water quality assessment indices can be selected to determine the level of water quality. In conclusion, this methodology provides a comprehensive framework for the evaluation of exposure risk to heavy metals, which can aid in the development of effective risk management strategies.

2.2 HEAVY METAL CONTAMINATION

Numerous scientific investigations have been dedicated to exploring the levels of heavy metal concentrations in dissolved substances and river sediments. Notable studies conducted by Swietlicka et al. (2017), Ozel et al. (2019), and Lu et al. (2019) have contributed valuable insights in this domain. In an effort to comprehend heavy metal behavior in aquatic environments, researchers have developed models based on geochemical processes. Significant contributions to this area of study have been made by Lindström and Håkanson (2001), De Blois et al. (2003), Braga et al. (2010), and Garneau et al. (2017).

Despite the progress made, several challenges persist when utilizing process-based models to simulate heavy metal behavior and concentrations in natural environments. Process-based models typically involve intricate descriptions of chemical processes that depend on numerous input variables. Consequently, the availability of monitoring data often falls short of the extensive data requirements of these models. Additionally, the complex internal mathematical representations of chemical processes may introduce uncertainties in parameter estimation and imprecise mathematical descriptions, negatively impacting the model's performance.

Moreover, the accuracy of model outcomes is highly sensitive to key parameters and can be hindered by insufficient validation data, resulting in reduced accuracy. Consequently, there is a clear need for the development of a relatively efficient and reliable model that can simulate heavy metal concentrations based on limited monitoring data.

In summary, the existing body of research highlights the requirement for a sophisticated model to accurately simulate heavy metal concentrations in aquatic environments, accounting for the intricate interplay of various geochemical processes. However, challenges such as limited data availability, parameter uncertainties, and inadequate validation hinder the accuracy and dependability of current process-based models. Consequently, it is crucial to develop a model capable of overcoming these challenges and providing rapid and accurate simulations of heavy metal behavior, particularly when confronted with limited monitoring data.

2.3 MACHINE LEARNING MODELS

In the past two decades, several models of artificial intelligence (AI) have been employed in heavy metal (HM) simulation. These models include artificial neural network (ANN), support vector machines (SVM), and response surface methodology (RSM), with the majority of studies utilizing ANN, SVM, and RSM for HM simulation. Ashrafi et al. (2019) and Dil et al. (2017) reported that the ANN model consistently outperformed multilinear regression (MLR) and RSM in predicting the removal percentage of Pb ions onto carboxylate-functionalized walnut shells (CFWS) adsorbent. The ANN model proved to be more effective than MLR and RSM in enhancing fitting in Pb prediction when scaled input data was used, and it took less time for prediction. However, the ANN model presents challenges, such as complexity and difficulty in establishing the cause-and-effect association between input and output variables.To improve the performance of the ANN model, researchers have employed various approaches. For instance, Gomez-Gonzalez et al. (2016) used a principal-component-analysis (PCA) featured selector with an ANN model, where the pattern-search approach was integrated with the LM algorithm to predict the adsorption capacity of Pb ions. In another study, González Costa et al. (2017) performed a comprehensive study to predict five HMs, including Pb sorption and retention, in soils using SVM, MLR, and regression-tree models. The authors found that noise in the MLR and regression-

tree models was minimized in the SVM model using the maximal parsimony principle. In addition, a group of scientists developed an SVM model with radial basis function (RBF), a kernel function such as the Gaussian, and grid-search methodology with tenfold cross-validation (CV) to predict Pb ion sorption (Parveen et al., 2017, 2016). The researchers validated the model against the MLR model and used Karush–Kuhn–Tucker conditions to minimize the issues of kernel function and biasedness.Wilson et al. (2013) focused on the ANN model with five responses, including Pb, where the furrier and breakthrough coefficients were utilized as the exoplanetary variables. The authors used the Bayesian regularization algorithm to normalize variables before training a single-layer ANN model to enhance its performance. However, the ANN model required a trial-and-error approach to design the best model topology, a finding also supported by Mandal et al. (2014). In conclusion, AI models have shown great potential in HM simulation, with ANN, SVM, and RSM being the most commonly used models. Although the ANN model has been found to be effective in predicting the removal percentage of Pb ions, it presents challenges such as complexity and difficulty in establishing the cause-and-effect association between input and output variables. To overcome these challenges, researchers have employed various approaches, including the use of SVM, MLR, and regression-tree models, as well as employing different algorithms and kernel functions. Despite the challenges posed by ANN models, they remain a promising tool in HM simulation.

2.4 INFERENCES

1) The existing body of research underscores the importance of investigating heavy metal concentrations in aquatic environments, specifically emphasizing the intricate interplay of geochemical processes. However, process-based models encounter challenges related to limited data availability, uncertain parameters, and inadequate validation, which hinder their accuracy and reliability.

2) Artificial intelligence (AI) models, including artificial neural networks (ANN), support vector machines (SVM), and response surface methodology (RSM), have emerged as promising approaches for simulating heavy metal behavior. ANN models, in particular, have demonstrated superiority in predicting the removal percentage of Pb ions compared to other methods like multilinear regression (MLR) and RSM, when input data is appropriately scaled.

3) Researchers have actively explored various strategies to enhance the performance of ANN models. For instance, they have incorporated feature selectors such as principal component analysis (PCA) and integrated optimization algorithms like the pattern-search approach with the LM algorithm to improve prediction accuracy.

4) SVM models, guided by the maximal parsimony principle, have shown efficacy in reducing noise and predicting the behavior of multiple heavy metals, including Pb sorption and retention in soils. SVM models have outperformed MLR and regression-tree models in noise reduction, making them suitable for capturing complex relationships within the data.

5) To address the complexity and challenge of establishing causal relationships in ANN models, researchers have investigated different approaches. They have proposed the utilization of various algorithms, kernel functions (e.g., Gaussian kernel), and validation techniques (e.g., grid-search methodology with tenfold cross-validation) to enhance the accuracy and reliability of ANN models in simulating heavy metal behavior.

2.5 SUMMARY

This chapter discussed the importance of understanding Based on these research findings, it is evident that the development of a sophisticated model capable of accurately simulating heavy metal concentrations in aquatic environments is crucial. Such a model should address the limitations associated with process-based models, harness the potential of AI models (particularly ANN and SVM), and account for challenges related to data availability, parameter uncertainties, and model complexity. Future research should focus on refining and advancing models that can overcome these challenges and provide rapid and reliable simulations, thereby facilitating a comprehensive understanding of heavy metal behavior and removal processes.

CHAPTER 3

METHODOLOGY

3.1 GENERAL

The previous chapter discussed the various kinds of literature focused on the present study. This chapter would discuss the various methodologies that will be adopted in the study to arrive at the objectives. The prediction of various heavy metals contamination in the middle ganga basin will be done using various machine learning models and performance evaluation will be carried out. The approach also intends to shed light on hyperparameter adjustment for enhanced model performance and to later calculate the water quality index using physiochemical parameters from the dataset that is obtained from feature importance plot.

3.2 STUDY AREA

The study was carried out in the Middle Gaga Basin lying in the Indian state of Uttar Pradesh. It is situated between 23° 52′ N and 31° 28′ N latitude and 77° 51′ E and 84° 38′ E longitude. The state's overall geographic area is 24,000 thousand hectares, or 7.33% of the total area of India, of which 16,573 thousand hectares are cultivated.

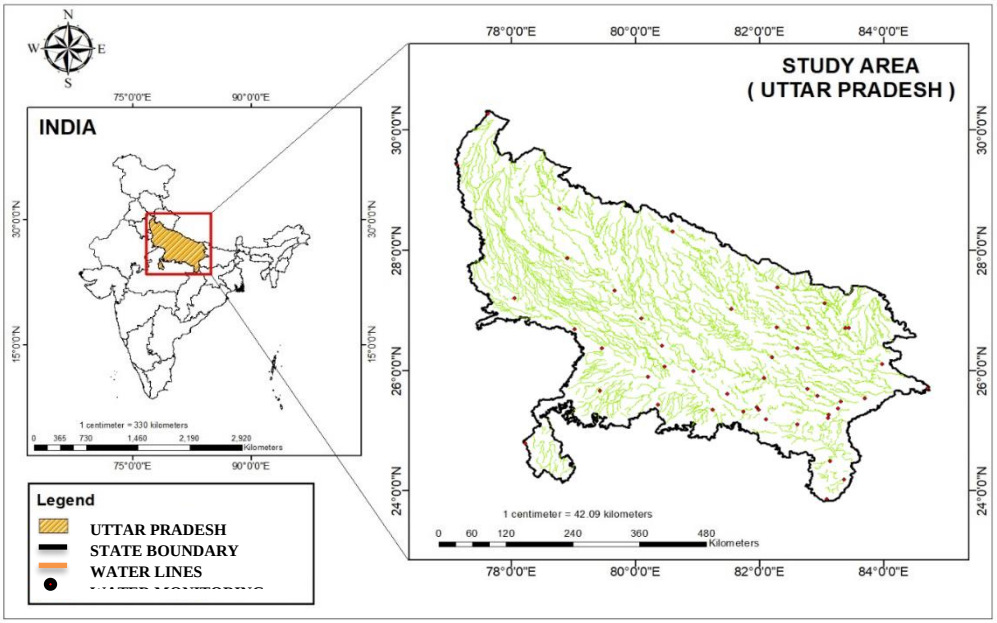


Fig. 3.1 Location of study area with water monitoring stations

A large portion of the state is located on the plains formed by the river Ganges and Yamuna. Fig. 3.1 shows the location map of study area and the water monitoring stations

situated near the water streams. These stations are considered the best reflecting points of heavy metal pollution caused by domestic and industrial activity as most of them are located near highly industrialized cities in the state.

3.3 Data Collection

For this study, we utilised physicochemical parameters data from the Central Water Commission (CWC). They maintain quality-controlled laboratory data, from a network of uniformly distributed across the basins all over the country. The monthly experimental water quality parameters data sets were collected for the middle ganga basin from 2007 to 2021. 43 different water quality monitoring stations spread across the lower part of the state are taken into consideration for this study.

Central Water Commission department collects water quality data comprising of 44 different physicochemical parameters which listed in the Table 3.1 below.

Table 3.1 Physicochemical Parameters present in the dataset.

Lead (mg/l)	Chemical oxygen demand
Alkalinity phenolphthalein (mg/l)	Dissolved oxygen (mg/l)
Total alkalinity	Alkalinity total
Arsenic (mg/l)	Dissolved oxygen saturation
Copper	Electrical conductivity field
Cadmium	Electrical conductivity (µmhos/cm)
Chromium (mg/l)	Fluoride
Nickel (mg/l)	Fecal coliforms (MPN/100ml)
Iron (mg/l)	Calcium hardness (mg/l)
Boron (mg/l)	Total hardness (mg/l)
Biochemical oxygen demand (mg/l)	Bicarbonate (mg/l)

Calcium (mg/l)	Potassium (mg/l)
Chloride	Magnesium (mg/l)
Carbonate (mg/l)	Sodium
Percent sodium	Secchi depth
Ammonia (mg/l)	Silicate (mg/l)
Nitrogen total oxidised	Sulphate (mg/l)
No^{3-}	Total coliforns (MPN/100ml)
Po^{4-}	Total dissolved solids (mg/l)
Ph	Temperature
Phosphorus total	Turbidity (NTU)
Sodium absorption ratio	

3.4 DATA PRE-PROCESSING

Data preprocessing is a crucial step that needs to be undertaken before incorporating the input data into a machine learning model. This step involves preparing and processing the data to ensure its suitability for modeling. Data pre-processing is a 4-stage process involving the following steps:

• Data integration

• Data cleaning and organizing

• Check for missing or null data and outliers

• Preparing training, testing, and validation data sets.

One common issue encountered in datasets is the presence of null or missing values, which must be addressed before proceeding with modelling. Handling null or missing

values is important as these gaps in the data can introduce bias and impact the accuracy of the model's predictions. There are different approaches to dealing with missing values, and one commonly used technique is either removing data points with missing values or replacing them with appropriate values. In the context of this study on heavy metal contamination prediction in the middle of the Ganga Basin, the null values were replaced with the mode value of the respective input feature. The methodology flow chart is shown in Fig. 3. 2.

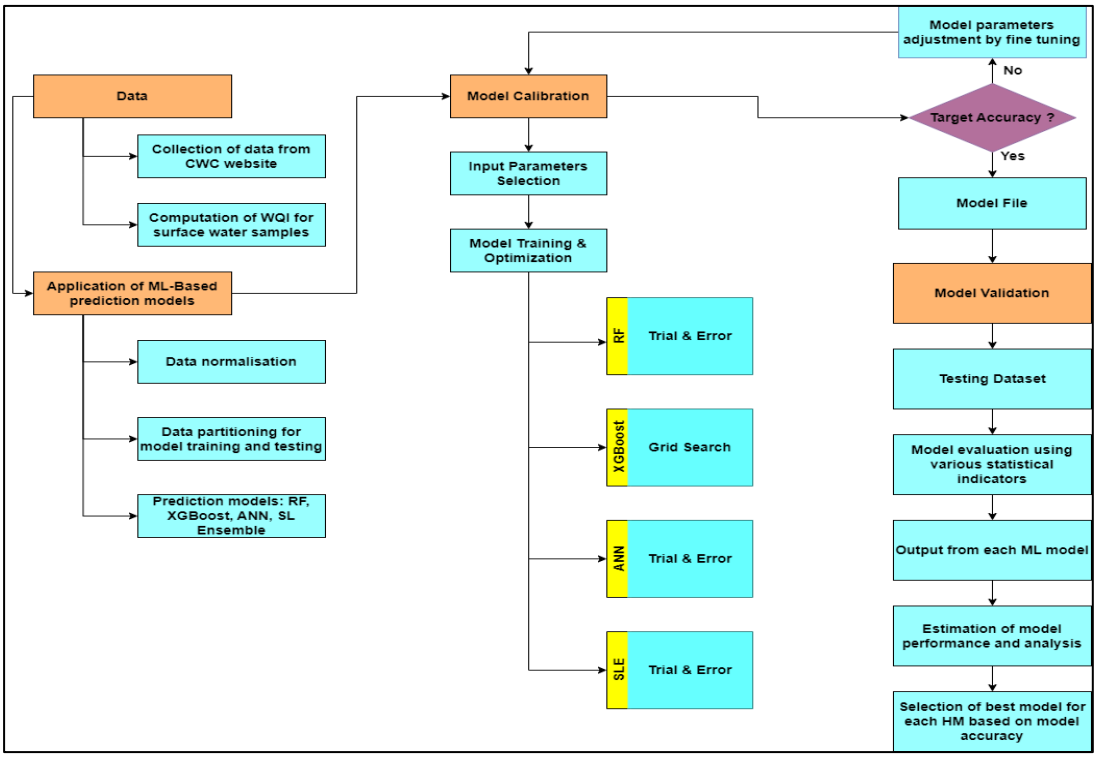


Fig. 3.2 Methodology flowchart adopted for the study

Another consideration in the data preprocessing phase is addressing outliers. Outliers are data points that significantly deviate from the majority of observations. They can arise due to various factors, such as measurement errors, extreme events with substantial variations in heavy metal concentrations, or the influence of nearby industrial activities and factories, which can lead to elevated levels of heavy metal concentrations.

Detecting and managing outliers is essential because they can distort the training process of the model and affect its overall performance. Outliers introduce noise and can influence the learned patterns of the model, resulting in less accurate predictions. Therefore, it is necessary to identify and remove outliers from the dataset before

proceeding with the modeling stage.

By effectively addressing missing values and outliers, it ensured the integrity and reliability of the dataset used for modeling heavy metal contamination in the middle of the Ganga Basin. This data preprocessing step played a crucial role in preparing a clean and dependable dataset, enabling the subsequent machine learning model to learn from the available information and make accurate predictions.

3.5 WQI ESTIMATION

The selection of significant water quality parameters is vital and key to having good representation of all indicators of water quality. Water quality parameters commonly used by various researchers include dissolved oxygen, total phosphates, temperature, pH, turbidity, chemical oxygen demand, fecal coliform, total solids, biochemical oxygen demand and nitrates. The weight associated with each parameter is based on its respective standards and the magnitude of the assigned weight indicates the parameter's significance and impact on the index. In this study, parameters used for the calculation of WQI are: BOD, pH, temperature, DO_{Sat} and TDS. Olubukola et al., (2021) suggested the weights for different parameters which are used for the calculation of WQI in this study is given below:

$$\frac{(0.19*BOD)+(0.12*pH)+(0.11*Temp)+(0.22*DO_{Sat})+(0.16*TDS)}{0.8}$$

Also, $$S_i = \frac{100}{(Ui - Li)} * (Ci - Li), \quad for\ U_i > C_i$$

$$S_i = \frac{100}{(Hi - Li)} * (H_i - Ci), \quad for\ U_i \leq C_i$$

where, U_i = Upper limit of the range for the parameter

L_i = Lower limit of the range for the parameter

H_i = Ideal value for the parameter

C_i = Observed concentration of the parameter

The WQI values for each of the stations were calculated and plotted on the map using Inverse Distance Weighted (IDW) toolbox in ArcGIS Software.

3.6 Applied Predictive Models

Machine learning and artificial intelligence have been in the spotlight recently with the development of the subject of data science, and several machine learning algorithms have either arisen or have grown in popularity. Based on the learning style that the algorithm has chosen, machine learning algorithms can be broadly classified into two types:

- Supervised Learning
- Unsupervised Learning

For this project only, Supervised learning following machine learning algorithm will be used. The multiple linear regression model, Decision tree model, Random Forest model, the Adaptive Boosting (AdaBoost) model, the Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN) which have been used in initial phase all belong to supervised learning.

3.6.1 Linear Regression

Multiple linear regression is a statistical method used to model the relationship between a dependent variable and two or more independent variables. It assumes a linear association between the dependent variable and the independent variables and aims to find the best-fit line that minimizes the difference between observed and predicted values.

To perform multiple linear regression, a dataset is collected with the dependent variable and multiple independent variables. Each independent variable represents a factor that may influence the dependent variable. The goal is to determine how changes in the independent variables affect the dependent variable.

In multiple linear regression, the model estimates coefficients for each independent variable, indicating the strength and direction of their impact on the dependent variable. These coefficients are calculated by minimizing the sum of squared differences between observed and predicted values, often using the ordinary least squares method. Once the coefficients are obtained, the model can be used to predict the dependent variable for new observations based on their corresponding independent variables.

Multiple linear regression finds applications in various fields such as economics, social sciences, finance, and marketing. It provides valuable insights into the relationships between variables and allows for the prediction of outcomes based on the values of independent variables. However, it is important to consider assumptions like linearity, independence, and normality for the model to produce accurate results.

3.6.2 Decision Tree

A decision tree is a widely used machine learning model that can be applied to both classification and regression problems. It adopts a flowchart-like structure where internal nodes represent features or attributes, branches denote decision rules, and leaf nodes represent outcomes or predicted values. Decision trees aim to create a predictive model based on input features. During the training phase, the decision tree algorithm recursively splits the data based on feature values. It selects the most informative feature that best divides the data into homogeneous subsets, employing measures like Gini impurity or information gain. This process continues until a stopping criterion is met, such as reaching a maximum depth or achieving a desired level of purity.

For prediction in a decision tree, input data traverses through the tree by following the path dictated by feature values. Ultimately, the data reaches a leaf node that provides the predicted class or value. Decision trees can effectively handle categorical and numerical features, making them versatile for different data types.

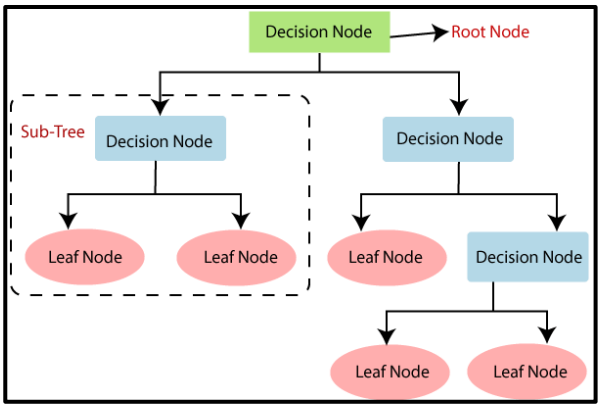


Fig. 3.3 Process of Decision Tree Model in Regression

Regarding regression, a decision tree can be used to predict a continuous target variable. Instead of predicting classes, the leaf nodes contain the predicted values. The decision tree algorithm partitions the data based on feature values, aiming to minimize variance

or the sum of squared differences between predicted and actual values. By following the appropriate path from the root to a leaf node, the decision tree can make regression predictions.

One advantage of decision trees is their interpretability and comprehensibility. They can capture intricate relationships between features and the target variable. However, decision trees are susceptible to overfitting, particularly when the tree becomes excessively deep and complex. Techniques like pruning, ensemble methods (e.g., random forests), and regularization can address overfitting and enhance the performance of decision tree models.

3.6.3 Random Forest

The random forest is a machine learning model widely used for both classification and regression tasks. It combines the concepts of decision trees and ensemble learning. Random forests consist of multiple decision trees, where each tree is built using a random subset of the training data and a random subset of features.

In the random forest algorithm, each decision tree is constructed independently. During training, a subset of the training data is randomly selected through bootstrapping, and at each node of the tree, a random subset of features is considered. This randomness introduces diversity among the trees, which helps reduce overfitting and improves the model's ability to generalize to unseen data.

To make predictions with a random forest, the input data is passed through each decision tree in the ensemble. For classification tasks, the random forest combines the predictions of individual trees and selects the class that receives the majority of votes as the final prediction. In regression tasks, the random forest aggregates the predicted values from the individual trees, such as by averaging or weighted averaging, to obtain the final regression prediction.

Random forests offer several advantages. They can handle high-dimensional datasets with numerous features and are less prone to overfitting compared to single decision trees. Random forests are robust to outliers and can capture complex relationships in the data. Additionally, they provide feature importance measures, which indicate the significance of each feature in the prediction process.

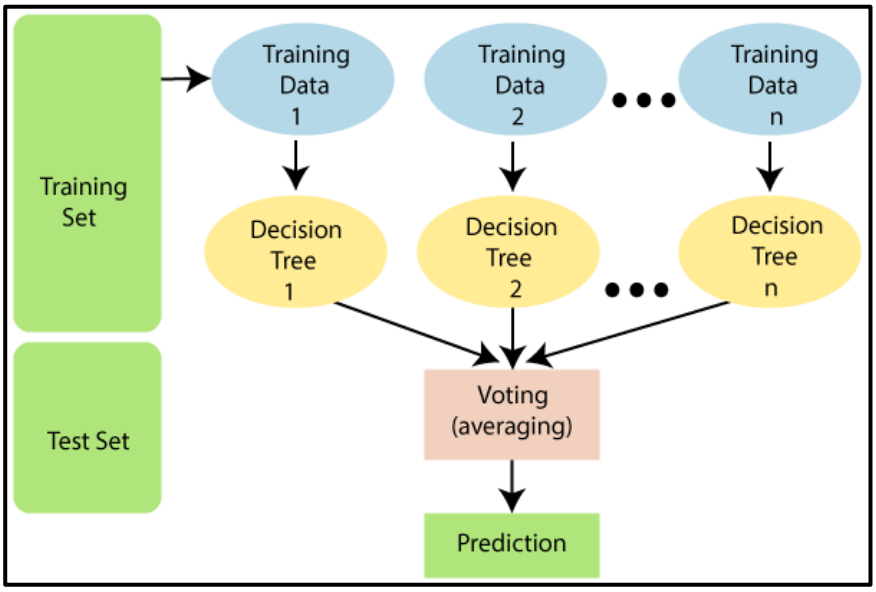


Fig. 3.4 Process of Random Forest Model in Regression

In regression problems, random forests are effective for predicting continuous target variables. Each decision tree in the random forest independently predicts a value, and the final regression prediction is obtained by aggregating the individual predictions. This aggregation can be done through averaging or weighted averaging. Random forests for regression provide accurate predictions, handle nonlinear relationships well, and are less sensitive to outliers compared to individual decision trees.

3.6.4 Adaptive Boosting

AdaBoost, short for Adaptive Boosting, is a machine learning model widely used for binary classification tasks. It excels at improving the performance of weak learners by combining them into a strong ensemble model. AdaBoost can also be extended to regression problems.

The AdaBoost algorithm iteratively trains a sequence of weak learners on different subsets of the training data. Each weak learner focuses on instances that were misclassified by previous learners, assigning them higher weights. This adaptive process ensures that subsequent learners pay more attention to challenging instances, effectively boosting their performance.

During training, AdaBoost assigns weights to each training sample, initially setting them equally. Weak learners are then trained on the weighted data, aiming to minimize the weighted classification error. After each iteration, the weights of misclassified

samples are increased, while the weights of correctly classified samples are decreased.

To make predictions, AdaBoost combines the outputs of all the weak learners. The final prediction is determined by weighing the votes of each learner based on its performance during training. Learners with higher accuracy have more influence on the final prediction.

Regarding regression, AdaBoost can be adapted to handle continuous target variables. Instead of using weak learners for classification, AdaBoost employs weak regressors. These regressors generate predictions based on weighted data samples and aim to minimize the weighted sum of squared errors.

AdaBoost offers several advantages. It is flexible and can use various weak learners, such as decision trees or simple rules. It performs well in practice and is less prone to overfitting. AdaBoost also provides an interpretable measure of feature importance, indicating which features contribute more to the model's predictions.

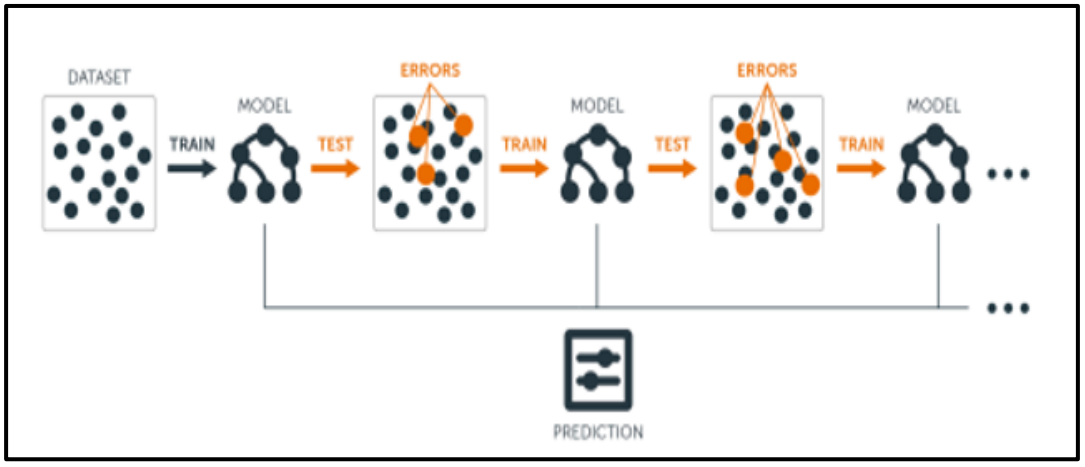


Fig. 3.5 Process of AdaBoost Model in Regression

In the context of regression tasks, AdaBoost combines weak regressors to form a powerful ensemble model. The weak regressors, trained on weighted data, contribute to the final regression prediction by minimizing the weighted sum of squared errors. AdaBoost regression effectively handles complex relationships between input features and the continuous target variable, delivering accurate predictions and robust performance.

3.6.5 Extreme grade boosting (XGBoost)

Extreme Gradient Boosting, commonly known as XGBoost, is a highly effective and efficient machine learning model used for both classification and regression tasks. It is based on the gradient boosting framework but is optimized to provide superior performance and speed.

XGBoost operates by iteratively building an ensemble of weak decision tree learners. Each tree is trained to correct the errors made by the previous trees in the ensemble. The model employs a gradient-based optimization technique, utilizing gradient descent to minimize a specified loss function such as mean squared error for regression or log loss for classification.

In XGBoost, decision trees are constructed in a depth-wise manner, dividing the data at each node based on the maximum reduction in the loss function. To prevent overfitting and enhance generalization, XGBoost applies regularization techniques such as shrinkage and column subsampling. It also incorporates a second-order approximation method known as "gradient boosting with second-order approximation" to further improve model performance.

In the context of regression, XGBoost aims to predict continuous target variables. It creates an ensemble of weak regression trees, with each tree learning to capture specific patterns and residuals in the data. The final regression prediction is obtained by aggregating the predictions from all the trees in the ensemble.

XGBoost offers several advantageous features as a regression model. It effectively handles complex relationships between features and the target variable, resulting in accurate predictions. The model automatically handles missing values in the data, which is a common challenge in real-world datasets. Additionally, XGBoost provides interpretable measures of feature importance, allowing for the identification of the most influential features in the regression process.

Overall, XGBoost is a highly regarded machine learning model due to its exceptional performance, speed, and versatility. Its ability to deliver accurate regression predictions has made it a preferred choice among data scientists and practitioners across various domains.

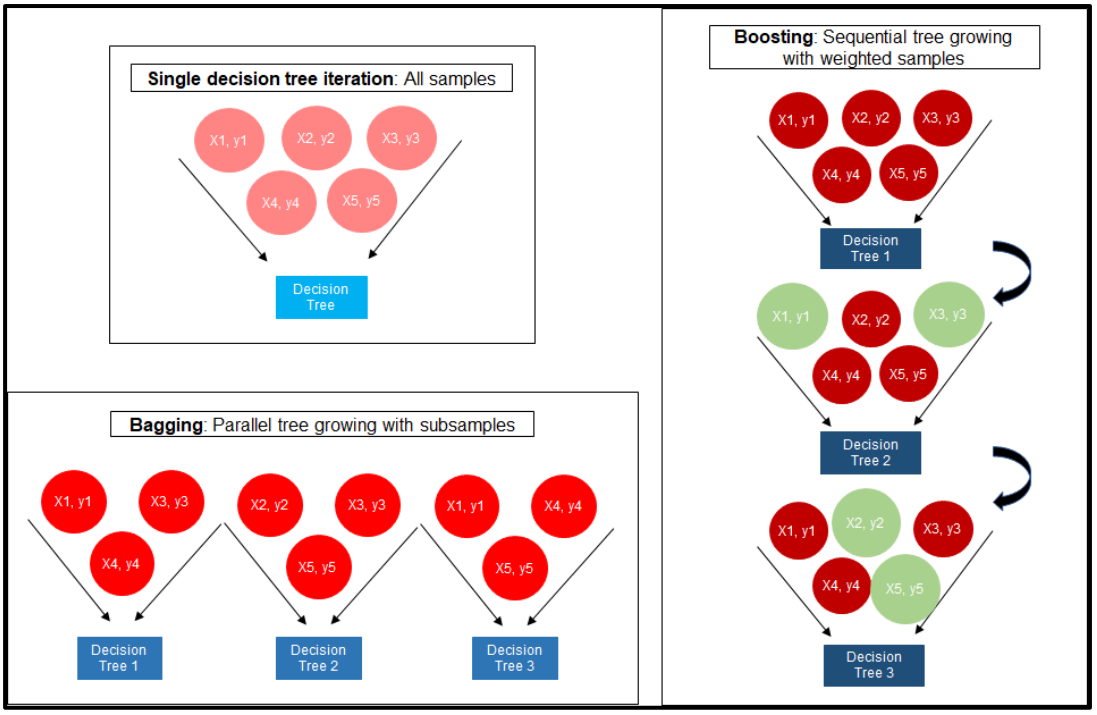


Fig. 3.6 Process of XGBoost Model in Regression

3.6.6 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functionality of the human brain. ANNs consist of interconnected nodes called neurons, organized in layers. Each neuron receives input signals, processes them, and generates an output signal that is passed on to other neurons. The connections between neurons have weights that determine the strength of the signal transmission. During the training phase, ANNs learn from labeled examples by adjusting the weights of these connections using optimization algorithms.

ANNs can be applied to various tasks, including regression. In an ANN regression model, the network learns to approximate a continuous target variable based on a set of input features. The input features are fed into the network, and they propagate forward through the layers, undergoing transformations through the weighted connections and activation functions. The final output of the network represents the predicted continuous value. The training process involves iteratively adjusting the weights to minimize the difference between the predicted values and the actual values in the training dataset. This allows the network to learn the underlying patterns and relationships in the data, enabling it to make accurate predictions on unseen inputs.

ANN regression models have proven to be effective in solving a wide range of regression problems, such as predicting housing prices, stock market trends, or weather forecasting.

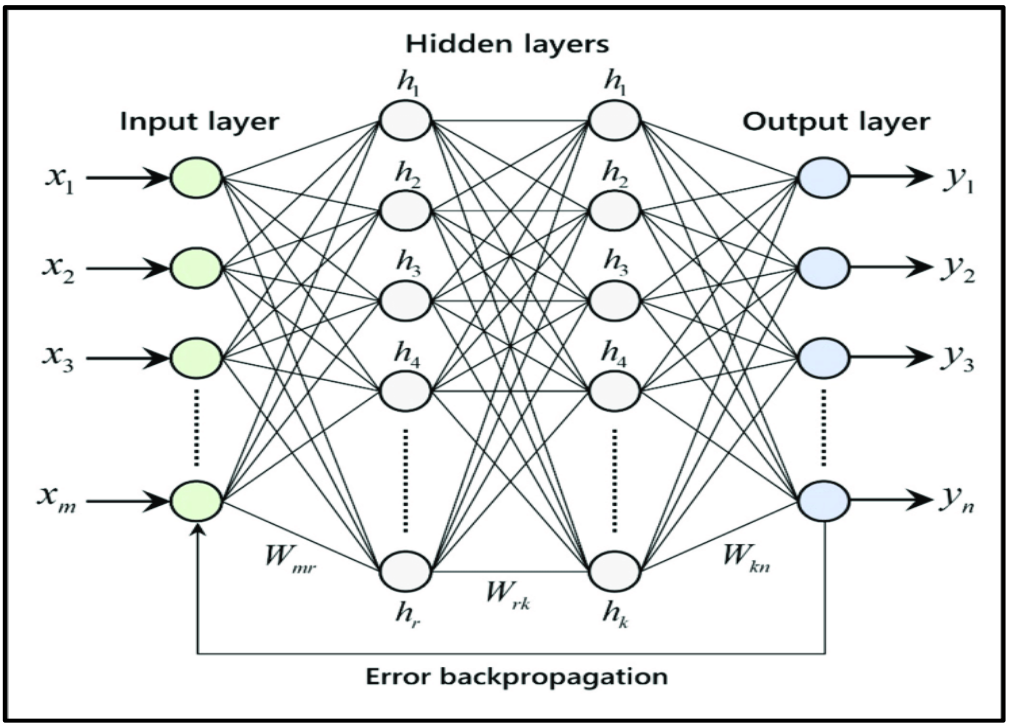


Fig. 3.7 Process of ANN Model

3.7 MODELING DEVELOPMENT AND PREDICTION METRICS

This research was established using various monthly experiment heavy metal assessments to predict Pb concentration. Also, it is utilized the constructed model for cross-station HM prediction assessment as per their best PC metrics.

PC was proposed by Karl Pearson in the 1880s, and it is recommended over Spearman's correlation coefficients in the field of science and engineering domain (Feng et al., 2019; Hauke and Kossowski, 2011). The correlation between the output variables and the selected predictors for each HM is calculated by Eq. 1

$$sim_i = \frac{\sum_{j=1}^{N}(X_j - \bar{X})(Y_{ij} - \bar{Y})}{(\sqrt{\sum_{j=1}^{N}(X_j - \bar{X})^2}) - (\sqrt{\sum_{j=1}^{N}(Y_{ij} - \bar{Y})^2})}$$

where sim_i is the 'similarity of the i-th project' of the historical project dataset, X_j and Y_{ij} which are the project-scale attribute data for the project to be reviewed and the i-th project, respectively. Also, X and Y are the average value of the project scale attribute data which expected to be reviewed and the i-th is the project of historical project data set.

The key point to enhance performance, reduce time consumption and bias during the modeling execution were achieved by the frequently used Eq. 2 (Sola and Sevilla, 1997; Tümer and Edebali, 2019).

$$X_{normal} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

where X_i are the original data, and X_{min} and X_{max} are minimal and maximal of the actual data, respectively. This normalization places all data within a uniform range, i.e., from 1 to 0, and consequently reduces noisiness, which leads to enhancing model performance.

Although there is no adequate procedure for the preparation of training and testing datasets in experiment and geographic cross-stations, it is recommended to divide the complete dataset into two subdatasets, training (70 %) and validation (30 %) (Bhagat et al., 2019).

3.7.1 OPTIMIZATION OF ML MODELS

In the case of ANN model prediction, its ability is enhanced by constructing two hidden layers (six and three neurons in the first and second hidden layers, respectively) to keep the learning process of the model running smoothly for thirteen input variables (Esmaeili and Aghababai Beni, 2015; Suditu et al., 2013b). The "resilient backpropagation with weight backtracking" algorithm was used with two repetitions for the neural network's training with linear output to enhance the mimicry phenomenon of the nonlinear data pattern. The model used the scaled dataset to enhance prediction performance (Yurtsever et al., 2014).

There is RF model that is built with 150 trees, eta (2), and mtry (2), which are

similar to the XGBoost parameters to reach that level of performance while keeping the model complexity at the same level as that in the candidate learner algorithm of the SL (Jiang et al., 2007; Ma et al., 2016; Sheridan, 2013). Moreover, maximal depth was a necessity to keep four, as two was not sufficient to enhance predictive performance.

3.7.2 PERFORMANCE METRICS (PMs)

The selection of specific efficiency criteria among several PMs, and the interpretation of the outcomes are challenging for the majority of researchers. Each PM sustains a different weight on the basis of several kinds of data simulations and actual behavior of the data. Statistical PMs of several applied AI learning models over the training and testing are in Tables 3 and 4. The computed performance metrics were error evaluators, i.e., root mean squared error (RMSE), mean absolute error (MAE), and efficiency evaluator, i.e., Nash–Sutcliffe efficiency (NSE). These PMs were applied to make the performance of the models reliable. R2 is a frequently used PM to evaluate predictive model performance as reported in the latest review research. This showed the degree of collinearity to achieve the efficiency and effectiveness of single HM prediction using AI models. Moreover, R2 also reported sensitiveness in the case of extreme values, and less sensitiveness to the proportional difference between predicted and real values (Bhagat et al., 2019; Chai and Draxler, 2014). MAPE seemed uncommon in the same research domain. It is used to estimate the predictiveness of flux prediction (Sekulić et al., 2019). Single- to hybrid-model performance is evaluated by MAPE. However, its sensitiveness due to high variation between predicted and actual value data is recommended to use along with other PMs (May Tzuc et al., 2018; Sonmez et al., 2018). RMSE is used to measure Pb sorption prediction by AI models (Parveen et al., 2016). RMSE is calculated the error between the model result and the response variable of the AI model. RMSE is used frequently in terms of a combination of HM studies. RMSE stands with its peripheral accuracy in line with the overall calculation process of the explicit model's performance. However, this value showed irregularities based on predictor type, such as a combination of HMs acting as a predictor, proportion-objected predictors, and fluctuating input variables. Using RMSE with additional errors measurement tools, such as an md metrics, makes results reliable (Ahmad and Haydar, 2016; Mendoza-Castillo et al., 2018; Moreno-Pérez et al., 2018). MAE metrics is the

obvious calculation in the case of Pb removal prediction study over the literature (May Tzuc et al., 2018). In general, the aim to calculate MAE is to compare the applied performance metrics and outcome of the applied AI models. Moreover, this is used to examine row error at the testing phase of hybrid models (Yaseen et al., 2016a). This also gives unbiased calculation in the case of data remaining without pre-processing such as cleaning, anomaly-scaled, and normalization (Kim and Kim, 2016; Shcherbakov et al., 2013).

The various model evaluation metrics used in the study are as follows:

Coefficient of determination (R^2): R^2 values provide a trustworthy indicator of model performance in most circumstances by demonstrating the degree of fit between the predicted and test variables.

$$R^2 = 1 - \frac{\sum(a_i - p_i)^2}{\sum(a_i - \mu_a)^2}$$

Mean Absolute error (MAE): Mean Absolute Error (MAE) is a measure of the average size of the mistakes in a collection of predictions, without taking their direction into account. It is measured as the average absolute difference between the predicted values and the actual values and is used to assess the effectiveness of a regression model

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i|$$

Root Mean Squared Error (RMSE): A model's short-term performance can be determined by comparing the actual difference between the estimated and measured values term by term using the RMSE.

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

Mean Squared Error (MSE): MSE tells you how close the points are to a regression line.

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

CHAPTER 4

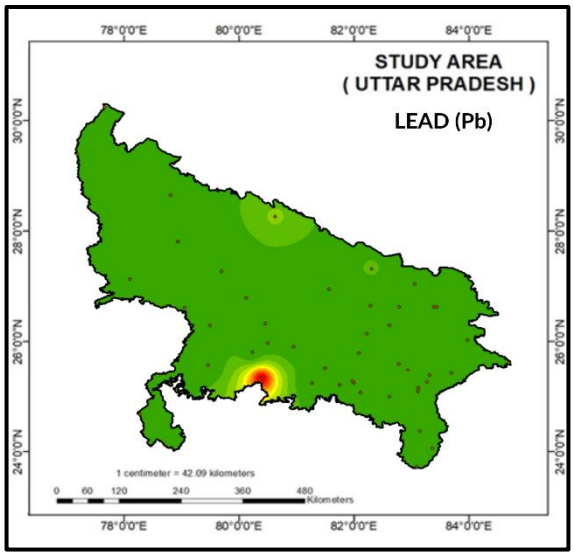
RESULTS AND DISCUSSION

4.1 GENERAL

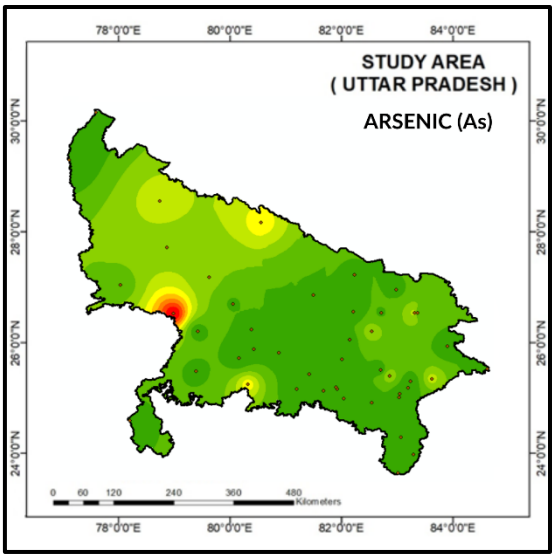
The chapter focuses on the performance and their evaluation of the models under different metrics. The chapter is divided into six sections, where each section investigates the performance and results of the respective model under consideration.

4.2 SPATIAL DISTRIBUTION OF WQI AND HM

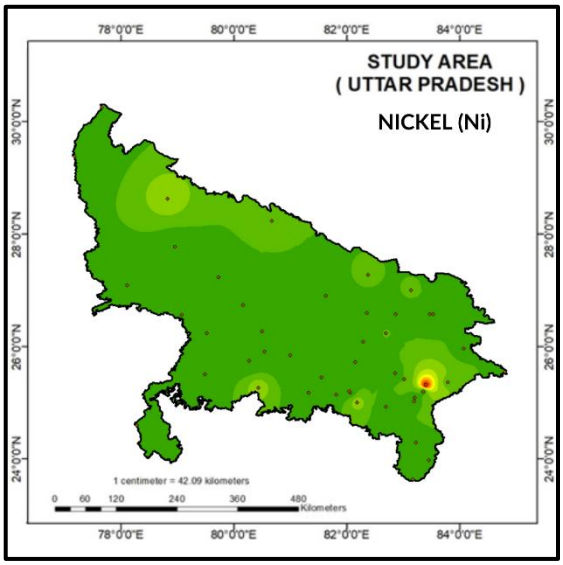
The WQI values for each of the stations were calculated and plotted on the map. The results are shown in Fig 3 below. Maps show that the hotspots for contamination are found to coincide with low WQI values, which highlights the importance of considering multiple factors when assessing water quality. Another point to be noted is that some stations with high contamination showed good WQI values. These results suggest that relying solely on heavy metal concentrations or WQI values alone may not provide a complete picture of water quality in the area.



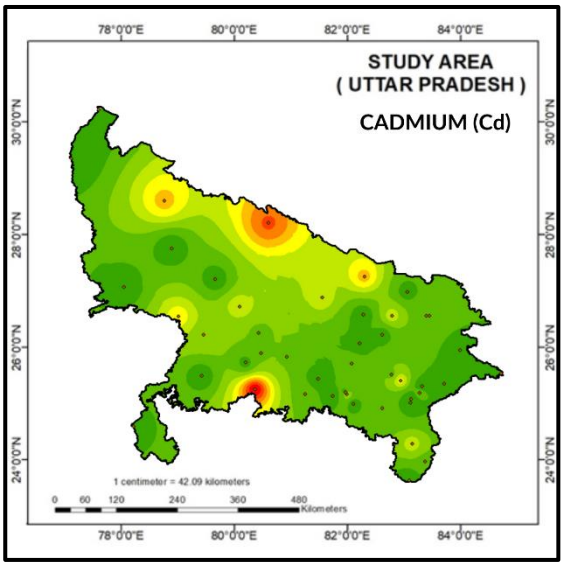
(a)



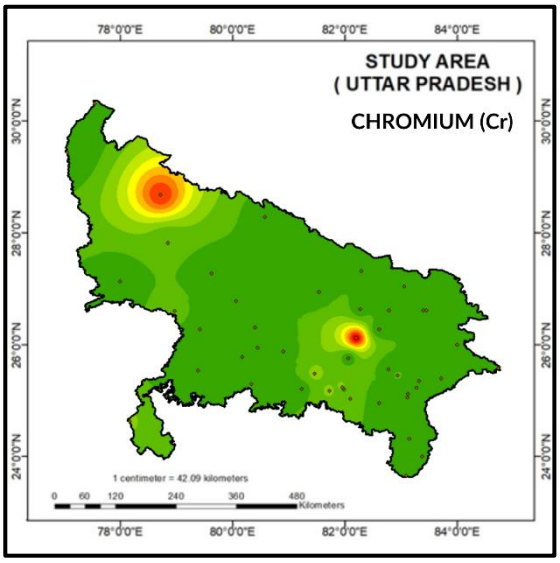
(b)



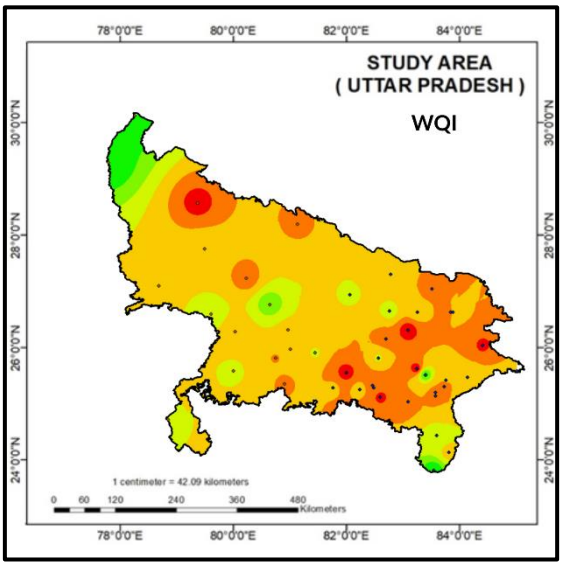
(c)



(d)



(e)



(f)

Fig. 4.1 Spatial Distribution maps of WQI and HMs concentration

4.3 CORRELATION ANALYSIS

To comprehend the relationship between the parameters of the input and the output, a correlation analysis of the data was conducted. Both the heavy metals and the physicochemical parameters were analyzed. The specifics of the physicochemical parameters employed in the investigation for each heavy metal are shown in Table 4.1.

Each model was run using 10 sets of physicochemical variables as input, and heavy metals as an output. These 10 sets were selected out of 38 parameters excluding 6 HMs as per PC analysis. Only those variables having PC coefficient values higher than 0.7 were taken into consideration.

Table 4.1 Input parameters taken from PC matrix

Pb	Cd	Cr	Ni	As
Secchi depth	Secchi depth	Secchi depth	Calcium	Secchi depth
pH	pH	Phosphate	pH	Potassium
BOD	Magnesium	BOD	Chloride	Magnesium
Dissolved Oxygen Saturation	Dissolved Oxygen Saturation	Dissolved Oxygen Saturation	Total Phosphrous	COD
Total Alkalinity	Temperature	Calcium	Total Alkalinity	Percent Sodium
TDS	TDS	Temperature	Sodium Adsorption rate	Silicate
Copper	Potassium	Silicate	Nitrite	Turbidity
Alkalinity	Silicate	Total Hardness	Percent Sodium	Nitrite
COD	COD	Fluoride	Total Hardness	Total Hardness
Turbidity	pH	COD	TDS	
Total Hardness	Silicate	Sodium Adsorption	Temperature	Dissolved Oxygen Saturation
Electrical Conductivity	Percent Sodium	Bicarbonate	Bicarbonate	TDS

4.4 ML MODELS

Various machine learning models have been used for the prediction of Heavy Metals. The models that are utilized for AQI modelling are: Decision Tree, Random Forest, XGBoost and ANN models.

4.4.1 Decision Tree (DT)

Decision tree gave the mixed results for the Heavy metals. While it was best performer for the prediction Cd with a R^2 score of 0.909, it was the worst performer for Ni having

a R^2 score of 0.665 only. The results obtained are mentioned in a tabular column in table

Table 4.2 Performance Metrices for Decision Tree

Heavy Metal	R^2 Score	MAE	MSE	RMSE
Lead	0.75477	0.098	0.1765	0.420119031
Arsenic	0.81135	0.087	0.09873	0.314213303
Cadmium	0.90856	1.02	0.09132	0.302191992
Chromium	0.75054	0.085	0.11268	0.335678418
Nickel	0.6648	0.092	0.11116	0.333406659

4.4.2 Random Forest model (RF model)

Like the Decision Tree, Random Forest also gave results with moderate performance. I came out best for predicting Cd while Cr came out at the lower end for its prediction. The results obtained are mentioned in a tabular column in table:

Table 4.3 Performance Metrices for Random Forest

Heavy Metal	R^2 Score	MAE	MSE	RMSE
Lead	0.788291	0.074	0.0123	0.110905365
Arsenic	0.830477	0.068	0.0098	0.098994949
Cadmium	0.861021	0.069	0.0099	0.099498744
Chromium	0.687945	0.085	0.0165	0.128452326
Nickel	0.691541	0.077	0.0143	0.119582607

4.4.3 Extreme grade boosting model (XGBoost model)

XGBoost came out as the best performing machine learning model for all the HMs. This model attained close to the ideal i.e. near the observed points, which were about 0.99. The results obtained are mentioned in a tabular column in table.

Table 4.4 Performance Metrices for XGBoost

Heavy Metal	R^2 Score	MAE	MSE	RMSE
Lead	0.999903	0.025	0.0075	0.08660254
Arsenic	0.999938	0.044	0.0068	0.082462113
Cadmium	0.999957	0.031	0.0077	0.087749644
Chromium	0.999784	0.039	0.0087	0.093273791
Nickel	0.999606	0.047	0.0076	0.087177979

4.4.4 Artificial Neural Network model (ANN)

ANN did not perform very well as evident from its performance metrics. It was the lowest performing model among all the other models deployed. The reason for the low performance could be the small dataset available which can cause overfitting of the model. The results obtained are tabulated in table

Table 4.5 Performance Metrices for ANN

Heavy Metal	R^2 Score	MAE	MSE	RMSE
Lead	0.529	2.54	3.129361	1.769
Arsenic	0.531	4.85	1.147041	1.071
Cadmium	0.532	3.33	3.500641	1.871
Chromium	0.528	8.98	1.157776	1.076
Nickel	0.544	4.68	1.993744	1.412

4.5 FEATURE IMPORTANCE

Feature importance is a crucial concept in machine learning that helps us understand the significance and contribution of each feature in a model. It enhances interpretability by revealing the relative influence of different features on the model's decisions. This knowledge aids in explaining and justifying the model's outcomes, especially when communicating with non-technical stakeholders. Additionally, feature importance assists in feature selection by identifying the most relevant features, simplifying the model, improving accuracy, and reducing computational costs.

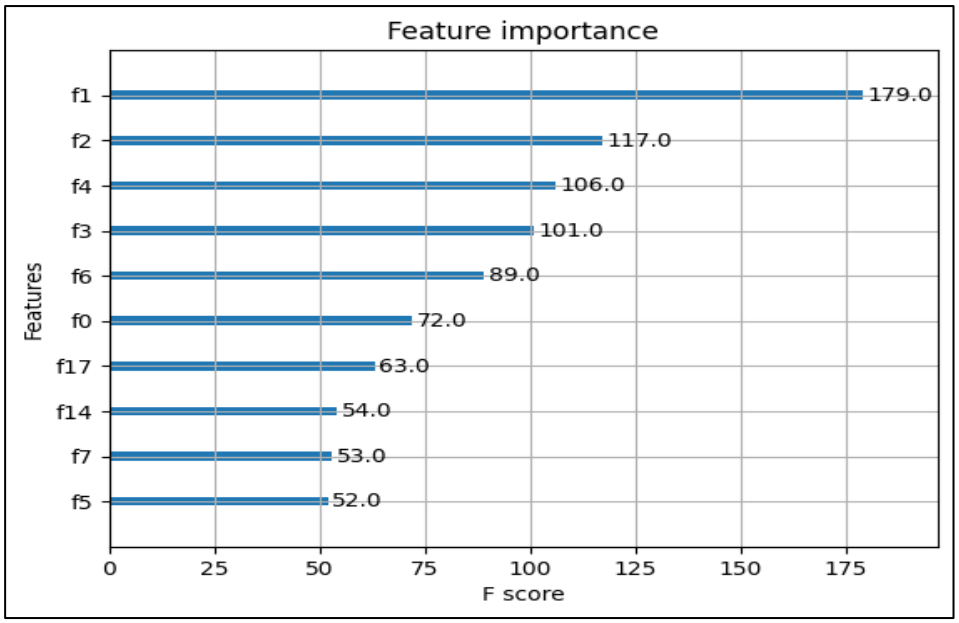


Fig. 4.2 Feature Importance Curve

Moreover, feature importance analysis helps identify redundant or irrelevant features, leading to more efficient models with reduced noise and overfitting risks.

Table 4.6 Most Important features for determination of each HM

Pb	Cd	Cr	Ni	As
Total Alkalinity	Temperature	Calcium	Total Alkalinity	Total Alkalinity
Copper	Secchi Depth	Iron	Sodium Absorption Rate	Potassium
Alkalinity Phenolphthalein	Silicate	Oxygen Demand	pH	Secchi Depth
Electrical Conductivity Field	Potassium	Fluoride	Total Hardness	Magnesium
Dissolved Oxygen		Total Hardness		

It also allows us to validate or challenge existing domain knowledge and assumptions, refining our understanding of the problem domain. Furthermore, feature importance is valuable for model debugging and performance improvement, helping identify data quality issues and guiding feature engineering efforts.

Lastly, feature importance analysis facilitates effective communication with

stakeholders by providing clear and understandable insights into the factors driving the model's predictions. This transparency fosters trust, collaboration, and comprehension among all involved parties. In summary, feature importance is a critical component of machine learning, enabling informed decision-making, model optimization, and establishing trust in the machine learning process.

4.6 COMPARISON OF MODEL PERFORMANCE

Overall statistical properties of the applied models revealed the superior predictability performed by the XGBoost model over the training and testing phase for all the HMs. Moreover, the other models remained closer to the XGBoost model in terms of R2 ; 15.5 %, 3.9 %, and 11.5 % less in the case of RF, SVM, and ANN, respectively. The XGBoost model attained close to the ideal i.e. near the observed points, which were about 0.99. e XGBoost model performance more consistent against ANN, SVM, and RF models, where the model had nonuniformity among the REs, but all models had their maximal error in the range of 10–20 except for the ANN model’s RE, unlike their minimal error, which had nonuniformity in all models.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 SUMMARY

The Middle Ganga Basin in India has been particularly affected by heavy metal pollution due to the discharge of untreated industrial wastewater into water sources and the major consequence is the degradation of the water quality over a period time. This study intends to analyse the effect of Heavy Metal on Water Quality and predict their concentration using Machine learning algorithms in Middle Ganga Basin encompassing Uttar Pradesh (UP) state, India. Evaluation of the HM concentration in the rivers is quite difficult and requires more time and effort. The study analyzed six heavy metals, including arsenic, cadmium, chromium, nickel, lead, and zinc, to identify the most significant predictors of heavy metal contamination. The study examined the spatial distribution of water quality index (WQI) and heavy metal concentrations using geographic information system (GIS) mapping, which revealed the relationship between hotspots of contamination and WQI in specific areas. Water quality data from 44 water quality monitoring stations in the basin, including physicochemical parameters and heavy metal concentrations, to train and evaluate the performance of four machine learning algorithms: Linear Regression, Random Forest, Artificial Neural Network, and Gradient Boosting Regression. The XGBoost model was found to be the best-performing algorithm, with an accuracy of over 98%, making it an effective tool for predicting heavy metal contamination in the region.

5.2 CONCLUSIONS

1. *Spatial Distribution Analysis:* Further research can delve into conducting a detailed spatial analysis of heavy metal (HM) contamination and its relationship with water quality index (WQI). By employing geospatial techniques such as geographic information systems (GIS) and remote sensing, researchers can generate high-resolution maps that illustrate the spatial distribution of HM concentrations and WQI values within the Middle Ganga Basin. This analysis would enable a comprehensive understanding of the hotspots of contamination and their corresponding impact on water quality. Moreover, it would provide valuable

insights into the spatial variability of WQI values in relation to HM contamination, allowing for targeted remediation efforts and resource allocation.

2. *Exploration of Machine Learning Techniques:* While XGBoost demonstrated robust predictive power for HM contamination, exploring other advanced ensemble techniques could further enhance the accuracy and robustness of the predictions. Techniques such as AdaBoost, Random Forest (RF) ensembles, or stacking can be investigated to leverage the strengths of multiple models and improve the overall predictive performance. By combining the outputs of various models, it is possible to capture diverse patterns and relationships in the data, thereby increasing the accuracy and generalizability of the predictions.

3. *Investigation of Model Limitations and Optimization:* To address the low performance of RF and artificial neural network (ANN) models, future research can focus on investigating the specific limitations and challenges associated with these algorithms. In the case of ANN, the reasons behind its lower predictive power, even after using scaled data, need to be thoroughly examined. Potential factors contributing to this could be insufficient data representation, inadequate model architecture, or issues related to hyperparameter tuning. By identifying and addressing these limitations, researchers can optimize the performance of RF and ANN models for HM contamination prediction.

5.3 FUTURE SCOPE

The study involving prediction of heavy metal contamination using machine learning algorithms in the middle ganga basin has several potential future scopes that can lead to further research in this field.

1) Transferability to Other Regions: The developed machine learning models and methodologies can be transferred and applied to other regions facing similar challenges of heavy metal contamination. Future research can focus on assessing the transferability and generalizability of the models to different geographical areas, considering regional variations in hydrology, land use, and pollutant sources.

2) Integration of Heavy metals Factors: WQI is influenced by various environmental factors such as land use, hydrological conditions, and climate change. Future studies

can explore the integration of HM into machine learning models to enhance prediction accuracy and reliability of these indices.

3) Long-term Monitoring and Analysis: Continuous monitoring of heavy metal concentrations and water quality parameters is crucial for understanding temporal variations and trends. Future research can emphasize long-term data collection and analysis to assess changes in heavy metal contamination over time. This would enable the identification of long-term patterns, the evaluation of mitigation measures, development of more robust models and the prediction of future contamination risks.

REFERENCES

1) **Alvarez-Iglesias, A., Hinde, J., Ferguson, J., Newell, J., 2017**. An alternative pruning-based approach to unbiased recursive partitioning. Compute. Stat. Data Anal. https:// doi.org/10.1016/j.csda.2016.08.011.

2) **Aryafar, A., Gholami, R., Rooki, R., Doulati Ardejani, F., 2012**. Heavy metal pollution assessment using support vector machine in the Shur River, Sarcheshmeh copper mine. Iran. Environ. Earth Sci. 67, 1191–1199. https://doi.org/10.1007/s12665-012- 1565-7.

3) **Ashrafi, M., Borzuie, H., Bagherian, G., Chamjangali, M.A., Nikoofard, H., 2019**. Artificial neural network and multiple linear regression for modeling sorption of Pb 2+ ions from aqueous solutions onto modified walnut shell. Separat. Sci. Technol. (Philadelphia) 1–12. https://doi.org/10.1080/01496395.2019.1577437.

4) **Bhagat, S. K., Tung, T. M., & Yaseen, Z. (2021)**. Heavy metal contamination prediction using ensemble model: Case study of Bay sedimentation, Australia. Journal of Hazardous Materials, 403, 123492. https://doi.org/10.1016/j.jhazmat.2020.123492

5) **Bui Quoc Lap, Thi-Thu-Hong Phan, Huu Du Nguyen, Le Xuan Quang, Phi Thi Hang, Nguyen Quang Phi, Vinh Truong Hoang, Pham Gia Linh, Bui Thi Thanh Hang**, Predicting Water Quality Index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system, Ecological Informatics, Volume 74, 2023, 101991, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2023.101991.

6) **Bisht, A.K., Singh R., Bhutiani, R., Bhatt, A. (2019)**. Artificial Neural Network Based Water Quality Forecasting Model for Ganga River. Volume-8 Issue-6, August 2019 https://doi.org/10.35940/ijeat.F8841.088619

7) **Cardwell, R.D., Deforest, D.K., Brix, K.V., Adams, W.J., 2013**. Do Cd, Cu, Ni, Pb, and Zn biomagnify in aquatic ecosystems, Rev. Environ. Contam. Toxicol. https://doi.org/ 10.1007/978-1-4614-6898-1_4.

8) **Chai, T., Draxler, R.R., 2014**. Root mean square error (RMSE) or mean absolute

error (MAE)? -Arguments against avoiding RMSE in the literature. Geosci. Model. Dev. 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

9) **Chapman, P.M., Wang, F., 2001**. Assessing sediment contamination in estuaries. Environ. Toxicol. Chem. https://doi.org/10.1002/etc.5620200102.

10) **Chen, X., Huang, L., Xie, D., Zhao, Q., 2018**. EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. Cell Death Dis. https://doi.org/ 10.1038/s41419-017-0003-x

11) **Enitan, I. T., Enitan, A. M., Odiyo, J. O. & Alhassan, M. M**. Human health risk assessment of trace metals in surface water due to leachate from the municipal dumpsite by pollution index: a case study from Ndawuse River, Abuja, Nigeria. Open Chem. 16, 214–227 (2018).

12) **Gomez-Gonzalez, R., Cerino-Córdova, F.J., Garcia-León, A.M., Soto-Regalado, E., DavilaGuzman, N.E., Salazar-Rabago, J.J., 2016**. Lead biosorption onto coffee grounds: comparative analysis of several optimization techniques using equilibrium adsorption models and ANN. J. Taiwan Inst. Chem. Eng. 68, 201–210. https://doi.org/10.1016/ j.jtice.2016.08.038.

13) **González Costa, J.J., Reigosa, M.J., Matías, J.M., Covelo, E.F., 2017**. Soil Cd, Cr, Cu, Ni, Pb and Zn sorption and retention models using SVM: variable selection and competitive model. Sci. Total Environ. 593–594, 508–522. https://doi.org/10.1016/j. scitotenv.2017.03.195.

14) **Krishnaraj, A., & Honnasiddaiah, R. (2022)**. Remote sensing and machine learning based framework for the assessment of spatio-temporal water quality in the Middle Ganga Basin. Environmental Science and Pollution Research, 29(43), 64939-64958.

15) **Küçükerdem TS, Kilit M, Saplıoğlu K (2019)** Determination of the number of clusters used in fuzzy inference systems by means of K-means and modeling of dam volume: Kestel Dam example. Pamukkale University Journal of Engineering Sciences 25(8):962–967

16) **Magdaleno A, De Cabo L, Arreghini S, Salinas S (2014)** Assessment of heavy

metal contamination and water quality in an urban river from. Argentina 18(1):113–120

17) **Ozel HU, Ozel HB, Cetin M, Sevik H, Gemici BT, Varol T (2019)** Base alteration of some heavy metal concentrations on local and seasonal in Bartin River. Environ Monit Assess 191(9):594 https://link. springer.com/article/10.1007/s10661-019-7753-0

18) **Taylor, K.E., 2001**. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106, 7183–7192. https://doi.org/10.1029/2000JD900719.

19) **Tessier, A., Campbell, P.G.C., Bisson, M., 1979**. Sequential extraction procedure for the speciation of particulate trace metals. Anal. Chem. https://doi.org/10.1021/ac50043a017.

20) **Trivedi, R. C. (2010)**. Water quality of the Ganga River–an overview. Aquatic Ecosystem Health & Management, 13(4), 347-351

21) **Ucun Ozel, H., Gemici, B. T., Gemici, E., Ozel, H. B., Cetin, M., & Sevik, H. (2020)**. Application of artificial neural networks to predict the heavy metal contamination in the Bartin River. Environmental Science and Pollution Research, 27, 42495-42512.

22) **WHO, 2017**. Safe Management of Wastes From Health-care Activities: a Summary. World Health Organization.