Journal of Mines, Metals and Fuels, *71(11): 1884-1895; 2023. DOI: 10.18311/jmmf/2023/35315*

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: *www.informaticsjournals.com/index.php/jmmf*

Classification and Regression-based Machine Learning Approach to Predict Mine Water Quality Index

Kulshresth Singh*, Dhananjay Kumar, Sudipta Mukhopadhyay and Indrajit Banerjee

Department of Mining Engineering, Indian Institute of Engineering Science and Technology, Shibpur, Howrah - 711103, West Bengal, India; kulshresthsingh.rs2019@mining.iiests.ac.in

Abstract

This work proposes a data mining-based prediction and development of the water quality index in mining areas. A mathematical equation for the index and predicted model is derived quantitatively in the study. Predicting water quality often involves applying conventional data mining techniques like classification and regression. Predictive learning and testing models can be evaluated using previous monitoring in real-time datasets and implementing k-fold cross-validation methods. The "decision trees" classification methodology outperforms other classification methods with 97.30% and 99.50% accuracy for training and testing model validation. MAE, RMSE, MSE, and R-squared are used in regression analysis to evaluate prediction accuracy and model performance. Regression model errors are absent with an R-squared value of 1. The present research showcases the efficacy of data mining techniques in accurately estimating mine water quality. These findings help improve mine water quality management.

Keywords: *Data Mining, Decision Trees Classification, Mine Water Quality Management, Predictive Modelling, Regression Analysis, Water Quality Index Prediction*

1.0 Introduction

Precipitation strength, rock reactive properties, porosity, permeability, water-rock contact duration, mining techniques, machinery, host rock composition, and local environmental factors affect mine water characteristics[1]. The Water Quality Index (WQI) helps choose the best treatment methods to satisfy quality standards by measuring water quality in one term. In the evaluation of water quality, parameters including Total Dissolved Solids levels (TDS), Electrical Conductivity (EC), pH levels, Dissolved Oxygen levels (DO), Biochemical Oxygen Demand (BOD), nitrates (NO_3), and Total Coliforms (TC) are considered. This comprehensive approach[2], utilizes a weighted arithmetic method for assessing water quality to classify water quality based on these critical factors. Exceeding these parameters' boundaries can harm health. The mine water quality index is a proven way to assess water quality in mining regions for human consumption.

1.1 Case Study

A significant study by MOIL (Manganese Ore India Limited) in Maharashtra focused on monitoring water quality across six specific groundwater and surface monitoring sites within the Kandri and Munsar mines. The study results indicated that, apart from surface water samples impacted by runoff, no contamination was identified in any of the samples. It is noteworthy to mention that the coliform values deviated from the norm. The water samples were found to be under the relevant

**Author for correspondence*

Indian standards, as indicated in the Executive Summary of the Environmental Impact Assessment/Environmental Management Plan for the Kandri Manganese Mine and the Executive Summary of the Environmental Impact Assessment/Environmental Management Plan for the Munsar Manganese Mine[3,4]. The research focuses on an estimated expanse of 2.5 km and furnishes comprehensive data. A thorough hydrogeological investigation was undertaken within the vicinity of the mining area. This investigation examined the extensive aquifer's daily variations and the potential consequences of the nearby manganese mine and soil erosion. The assessment of the impacts of the mines and roadways was conducted through the measurement of water levels in the pre-existing excavation wells. In addition, it is imperative to implement efficient control mechanisms to minimise the emission of airborne particles from the mining operation and the landfill. It underscores the need to implement appropriate mitigation solutions.

1.2 Standards of Water Quality in India

The effective utilization of water resources is contingent upon maintaining an optimal degree of purity that aligns with the planned usage of a given water body. The amount of purity required for water varies depending on its application, highlighting the significance of categorizing water consumption and setting quality standards accordingly[5]. Central Pollution Control Board of India has published detailed guidelines regarding the essential chemical characteristics of water, also known as fundamental water quality standards. Moreover, under the established norm IS 2296:1992, the Bureau of Indian Standards (BIS) has delineated water quality standards catering to various applications. Essential contributions to formulating stipulated quality criteria for water bodies are derived from the Hydrology and Water Resources Information System[6,7] in India. Comprehending and following these criteria is crucial for proficiently overseeing and preserving water quality to fulfill the distinct requirements of diverse applications.

1.3 Relevant Paper Studies

Past research shows that data mining techniques are essential in the environmental industry.

Rajagopalan and Lall simulated average precipitation and environmental conditions using K.N.N[9]. Bressler *et al.*, developed reservoir system operating rules using decision trees[10].

Using several methodologies, Hyvonen *et al.*, classified essential elements in atmospheric aerosol particle generation[11]. Palani *et al.* (2008) use an artificial neural network to predict seawater temperature[12]. Mucherino *et al.*, tested K-nearest neighbor, ANN, and SVM techniques for agricultural difficulties[13].

Gibert *et al.*, categorised complicated wastewater treatment facility trends using library knowledge discovery[14]. Gazzaz *et al.*, as well as Motamarri and Boccelli, utilised Artificial Neural Networks (ANN) as a methodology for the classification of water quality[15,16]. Radojevic *et al.*, used decision trees and clustering analysis to study the water in reservoirs with coliform microbes and established parameters[17].

Table 1. Features and optimal usage of water quality standards in India[8]

Parameters	Recommended Usage				
	A	B	C	D	E
TDS (mg/l)	500	-	1500	-	2100
EC (micromhos/cm)	-	-	-	-	2250
pH level	6.5 – 8.5	6.5 – 8.5	6.0 – 9.0	6.5 – 8.5	6.0 – 8.5
DO (mg/l)	6	5	4	4	-
BOD (mg/l)	2	3	3	-	-
NO_3 (mg/l)	20	-	50	-	-
TC (MPN/100 ml)	50	500	5000	-	-

Verma *et al*., developed statistical prediction models for daily wastewater total suspended particle fluctuations using data mining[18]. Kovcs *et al*., made a good case for using clustering and selected analysis to get homogeneous qualitative water samples from Neusiedler Bay, Europe's most famous and westernmost steppe lake[19].

Liu and Lu examined ANN and SVM methods for forecasting Total Nitrogen (TN) and Total Phosphorus (TP) in nonpoint agricultural runoff-affected waterways[20]. Mohammadpour *et al*., predicted built-in wetland water quality using SVM and ANN[21].

These studies demonstrate data mining's efficacy and broad environmental applications, guiding environmental management and decision-making.

1.4 Proposed Approach

The proposed approach in this study aims to predict and develop a water quality index in the mining sector using data mining techniques. These techniques are precious for dealing with heterogeneity, large datasets, data inconsistency, and missing data commonly encountered in water quality data collection at various mining sites. Data mining is fast and effective in identifying patterns in datasets and classifying data based on those patterns, providing decision-makers with valuable insights[22]. The process for forecasting and constructing the water quality index encompasses a series of stages, as seen in Figure 1.

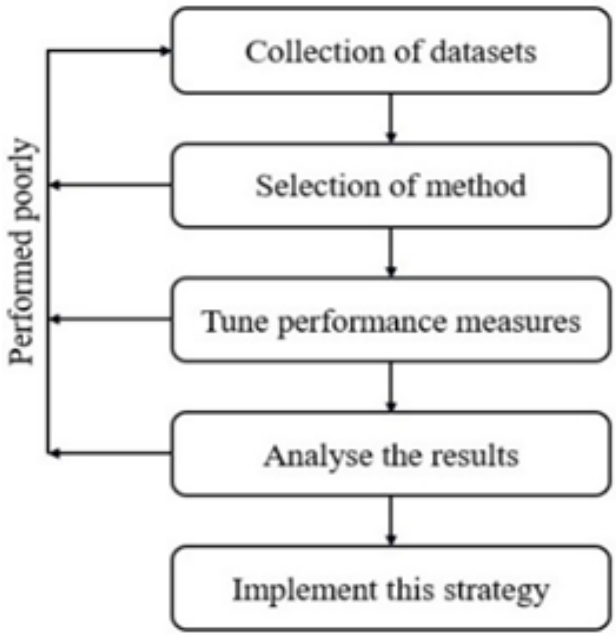

Figure 1. Typical procedures for supervised learning.

The process of data collection involves obtaining information from cloud server hosts and pertinent regulatory entities such as State Pollution Control Boards (SPCBs) or Pollution Control Committees (PCCs), under the National Water Monitoring Programme (NWMP) guidelines[23] outlined in the report titled "Water Quality of Medium and Minor Rivers, 2019." The computation of the water quality index in mine environments involves the application of the "Weighted Arithmetic Water Quality Index" methodology for index calculation. Value assignment systematically involves the allocation of values, taking into consideration the water quality index grade. This approach ensures a consistent classification of values. This study explores the application of supervised machine learning techniques in predicting datasets. The research involves the implementation of a supervised machine learning methodology that incorporates classification and regression approaches. The evaluation of prediction models involves the assessment of their performance using regression measures, which are evaluated by mathematical expressions.

This study utilises data mining methodologies to evaluate the water quality index throughout the mining industry and extract significant insights to facilitate informed decision-making. Two supervised machine learning methodologies are utilised to develop prediction models:

1. The classification methodology[11] is specifically tailored for dependent variables with a limited set of non-sequential values. In this context, the accuracy of predictions is evaluated based on the occurrence of misclassification.
2. In machine learning, regression is a statistical strategy that calculates the squared difference between observed and predicted values. This method is advantageous when dealing with dependent variables that exhibit varying levels of constancy or order[11].

This work endeavors to utilize data mining techniques and machine learning methodologies to construct precise prediction models for the water quality index in the mining sector. The objective is to provide significant insights that can inform decision-making processes.

2. Materials and Methods

2.1 Collecting Datasets

During the preliminary phase of this study, data samples were gathered from the mining sector. The data gathering procedure includes using a cloud server, a virtual server hosted and distributed across the internet via a cloud computing platform. This cloud server offers on-demand services to customers[24]. According to Kapil *et al*., Cloud computing lets users store and retrieve data via servers offered by cloud service providers, with data storage being a fundamental service within this framework[25].

This study gathered data on seven frequently employed water quality indicators. The parameters mentioned earlier include Total Dissolved Solids (TDS), Electrical Conductivity (EC), pH levels, Dissolved Oxygen levels (DO), Biochemical Oxygen Demand (BOD), Nitrates (NO3), and Total Coliforms (TC). The pH and TDS values were obtained using a cloud-based system that archives live monitoring data from the Kandri and Munsar manganese mines (MOIL). The device was purposefully planned, developed, and implemented in a research study under the "Real-Time Water Quality Monitoring System" project.

To augment the dataset, the additional water quality indicators, specifically EC, DO, NO_3, BOD, and TC, were obtained from the "Water Quality of Medium and Minor River 2019" report, which was compiled by the State Pollution Control Board (SPCB) as part of the National Water Quality Monitoring Programme (NWMP) overseen by the Pollution Control Committee (PCC). The water quality index was anticipated and calculated using supplementary metrics.

The procedure of data collecting entailed the meticulous selection of pertinent measurements from various sources, intending to ensure a thorough depiction of the water quality factors for the research investigation.

2.2 Mathematical Approach

The research project utilises a mathematical methodology centred on developing and progressing a water quality monitoring programme[26]. This programme evaluates the physical, chemical, and biological elements influencing water quality in different sources and places. Monitoring these parameters is of utmost importance since they directly influence human health, and levels above established thresholds can have adverse effects. The study used the widely acknowledged Water Quality Index (WQI) to quantify water quality, which serves as a practical methodology for assessing the suitability of water resources for human usage. The importance of the Water Quality Index (WQI) has been recognised by several reputable sources, such as the relevance of the Research and Development (R&D) programmes in the Ministry of Drinking Water and Sanitation (MOIL) in 2021[27], the Guidelines for drinking water quality in 2012[28], the Bureau of Indian Standards in 2012[28], and the United States Environmental Protection Agency (EPA) in 2009[5].

The research uses the "Weighted Arithmetic Water Quality Index" methodology with other approaches for assessing water quality. This technique was selected based on its appropriateness for the current inquiry. This approach enables the inclusion of weighted elements that accurately represent the varying significance of distinct water quality criteria within the comprehensive evaluation.

The state of the water monitoring programme will undergo ongoing modifications and updates during the study project. Adaptability is crucial in acquiring new information and data, expanding understanding, and advancing through the many phases of a mine's life cycle. Through the use of the selected mathematical methodology, this study seeks to thoroughly assess the water quality of the mine and provide valuable contributions to enhancing decision-making and management strategies within this field.

2.3 Data Mining Techniques Approach

The research study included modelling techniques employing one-year historical records acquired from mines via cloud service provider hosts. The dataset was randomly partitioned into comparable k-disjoint sets to create a learning and testing framework using recurrent k-fold cross-validation. Each set exhibits a homogeneous class distribution. The datasets are subjected to training and testing procedures employing a machine learning technique for classification and regression tasks.

The research study employed several classification and regression learning strategies within the supervised machine learning framework.

Classification learning techniques are a fundamental aspect of machine learning. These techniques include categorising data into distinct classes or categories

based on certain features or attributes. By employing classification learning techniques, researchers and practitioners can develop.

In this current research study, a supervised machine learning framework was utilised to explore several classification and regression learning approaches.

A. Classification Learning Techniques

- Decision Trees
- Naïve Bayes
- Support Vector Machine
- Ensemble classifier
- Neural Network
- Kernel Approximation

B. Regression Learning Techniques

- Linear Regression Model
- Gaussian Process Regression Model

Each strategy employs distinct approaches to the fundamental relationship structure between the indicator parameters and the class label, resulting in possible variances in their performance when applied to the same dataset. The classifier performance evaluation on unfamiliar datasets relies on diverse metrics data mining techniques offer. The present study utilises the ACCURACY technique to evaluate the performance of each classifier, adhering to the standards stated by Ali *et al*., and Melo, as specified in the Guide Manual: Water and Waste Water published by the Central Pollution Control Board[29-31].

This research study uses various data mining approaches to examine the correlation between indicator parameters and the mine water quality index. The objective is to conduct a comprehensive investigation of the predicting capacities of each classifier.

2.4 Proposed Data Analysis Techniques

The Water Quality Index is derived using a mathematical methodology called the Weighted Arithmetic Water Quality Index Method. The scientific community has widely adopted it, with many academics using it in their work[32-34]. The following equation determines the Water Quality Index (WQI)[35]:

$$WQI = \frac{\sum Q_i W_i}{\sum W_i} \quad (1)$$

Qi, the quality rating scale, is determined by the following equation, which is applied to each parameter:

$$Q_i = 100\left[\frac{(V_i - V_o)}{(S_i - V_o)}\right] \quad (2)$$

V_i represents the estimated concentration of the i^{th} parameter in the analysed water. The optimal value of this parameter in pure water is denoted as V_o. The initial value (V_o) is zero, except for a pH value of 7.0 and a Dissolved Oxygen (DO) concentration of 14.6 mg/l. The symbol "S_i" represents the suggested standard value for the ith parameter.

The below equation is utilised to determine the unit weight (W_i) assigned to each water quality metric:

$$W_i = \frac{k}{S_i} \quad (3)$$

Where k is a proportional constant which can be calculated using the following formula:

$$k = \frac{k}{\sum\left(\frac{1}{S_i}\right)} \quad (4)$$

Table 2. The assessment of the Water Quality Index (WQI) value

WQI Range	Water Quality Rating	Quality Grade
0 – 25	Exceptional Purity	Grade A
26 – 50	High-Quality Standards	Grade B
51 – 75	Moderate Quality	Grade C
76 – 100	Below Standard Quality	Grade D
Above 100	Unfit for Drinking	Grade E

The water quality rating, as determined by the water quality index technique, is presented as follows:

The next part of this work uses supervised machine learning to create a prediction model. Machine learning algorithms classification and regression are used. The model is trained and tested using K-fold cross-validation. The large dataset is split into training and testing subsets to use supervised machine learning. Subsets are further separated into halves. First, categorisation learning is used to build the prediction model. The model is trained and tested using repeated K-fold cross-validation. This study uses 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 K-fold values.

The methodology employed to assess the precision of a prediction model is the numerical approach[36] given by:

$$Accuracy = \frac{Correct\ prediction}{prediction} \quad (5)$$

The accuracy of predictions for the test data is measured as a percentage[37]. The calculation may be readily performed by simply dividing the count of accurate forecasts by the count of all guesses.

This study builds a prediction model for real-time mine water quality forecasts and will be done through regression learning. The regression learning method used repeated K-fold cross-validation.

The predictor-responder link is evident in a data model. The optimum linear model parameters for a dataset are estimated using linear regression. The most common linear regression approach is least-squares, which fits lines and polynomials[38].

Linear regression shows how a variable called 'y' is connected to one or more variables called 'independent variables', x_1, x_2,..., x_n. Simple linear regression examines the connection between independent and dependent variables.

$$y = \beta_0 + \beta_1 x + \epsilon \quad (6)$$

In this context, β_0 represents the y-intercept, β_1 denotes the slope (sometimes referred to as the regression coefficient), and ϵ represents the error term.

Consider a collection of n observed values of x and y denoted as (x_1,y_1), (x_2,y_2), ..., (x_n,y_n). As mentioned, the values provide a set of linear equations when the basic linear regression model is employed. The equations, as mentioned earlier, are expressed in matrix notation.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad (7)$$

Let

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, B = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad (8)$$

The present association is denoted by the equation Y = XB. The coefficient matrix B in MATLAB may be computed using the 'mldivide' operator, represented as B = X\Y. In the context of a specific dataset, namely the "mine Water Quality Index (WQI)," it is possible to build a linear regression relationship denoted as $y = \beta_1 x$. This relationship may be determined by loading the WQI data into the variable 'y' and the water quality metrics data into the variable 'x' and utilizing the \ operator. This operator does a least squares regression analysis to ascertain the correlation between the Water Quality Index (WQI) and other indicators related to water quality.

2.5 Evaluation Metrics for Regression Models

MSE, MAE, RMSE, and R-squared are used in regression analysis to measure prediction error rates and model performance[39-43].

The Mean Absolute Error (MAE), obtained by taking the average difference between the real and expected values throughout the dataset, measures the discrepancy between these values (Equation (9)).

The Mean Squared Error (MSE) measures the discrepancy between the actual and predicted values. It is computed by taking the average of the squared differences throughout the whole dataset, as shown in Equation (10).

Equation (11) represents the Root Mean Squared Error (RMSE).

The coefficient of determination, commonly called R-squared, quantifies the degree of match between the observed values and the original data. The value is bounded between 0 and 1 and is denoted as a percentage. Equation (12) demonstrates a positive correlation between the quality of the model and its corresponding

value, indicating that as the model improves, the value also increases.

The key conclusion drawn from R^2 is that a higher value, specifically 1.0, is more desirable. It suggests that the regression model is devoid of errors.

A coefficient of determination (R^2) equal to zero signifies that the regression model does not provide any improvement over just using the mean value. It suggests the model does not use any information from the other variables.

A negative R^2 score suggests the observed results are poor and below average. However, the measure of summed squared error may not hold the utmost significance, which is deemed acceptable.

The mathematical formulas used to assess the performance of regression models are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}| \quad (9)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y})^2 \quad (10)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y})^2} \quad (11)$$

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y})^2}{\sum(Y_i - \bar{Y})^2} \quad (12)$$

Where,

$\hat{Y}$ represents the estimated or projected value of Y.

$\bar{Y}$ indicates the mean value of the variable Y

3.0 Results and Discussion

3.1 The Outcome of Mathematical Analysis

This paper comprehensively elucidates the computational approach to determining the overall Water Quality Index (WQI). As mentioned earlier, the computation is utilised to analyse all datasets on the water quality parameters investigated within the scope of this study. The methodology section provides a comprehensive explanation of the computation process. Figure 2 depicts a visual depiction of the computed Water Quality Index (WQI) value for a specific year in the context of mining activities. The graph visually represents the temporal fluctuations in the mine's Water Quality Index (WQI),

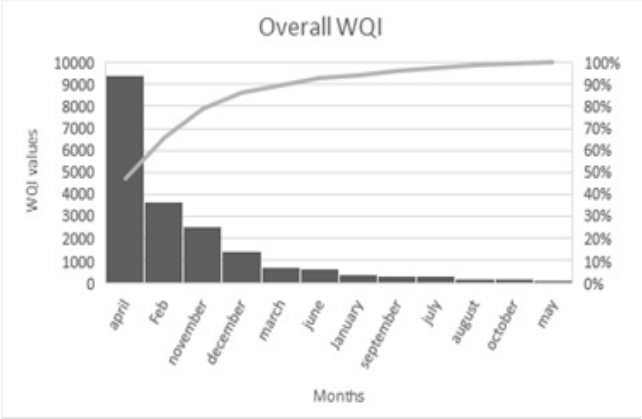

Figure 2. Mine-wide WQI value throughout the year is shown graphically.

illustrating any discernible trends or patterns throughout the year. The x-axis represents the horizontal axis. It represents the period, while the vertical axis is designated as the y-axis and represents the mine Water Quality Index (WQI) value. The graph depicts the mine's Water Quality Index (WQI) variability and offers insights into the broader water quality patterns.

3.2 Results from the Data Mining Techniques Approach

The Results section of this research study focuses on using supervised machine learning to build a model for predicting outcomes. The methodology involved the utilization of two machine-learning algorithms, namely classification, and regression, inside a supervised

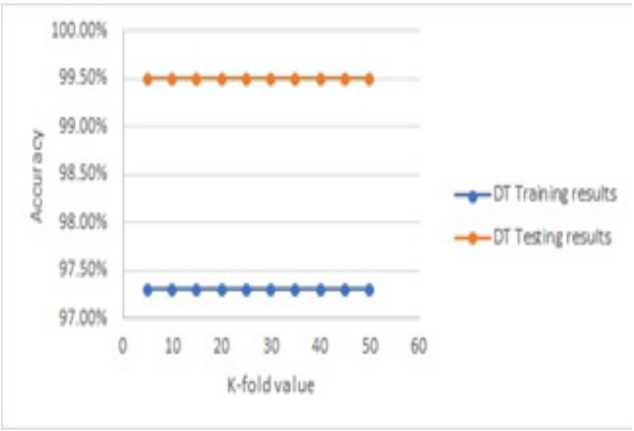

Figure 3. Decision tree accuracy results.

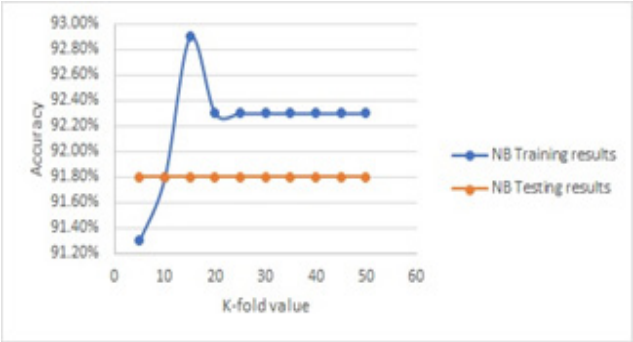

Figure 4. The accuracy of naive Bayes.

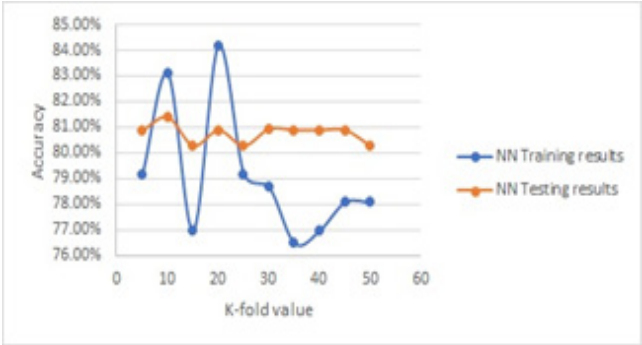

Figure 7. Results of the neural network's precision.

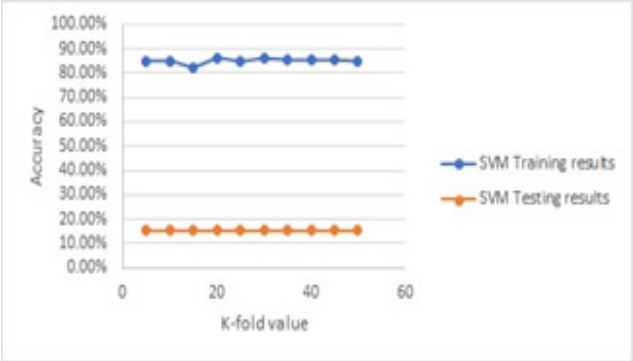

Figure 5. Support vector machine accuracy results.

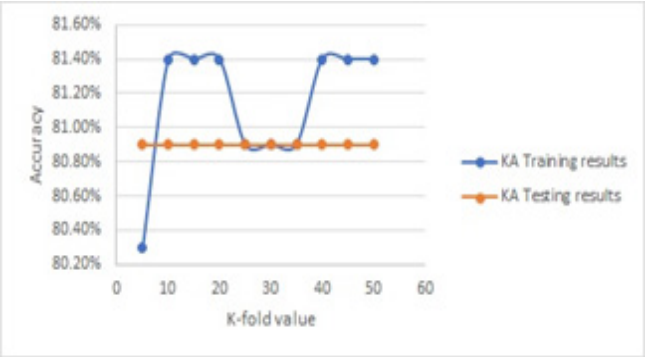

Figure 8. Results of Kernel Approximation's precision.

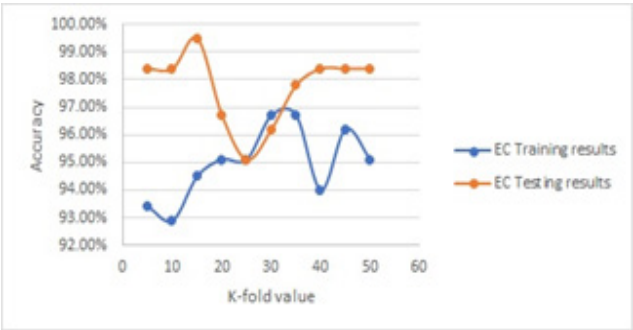

Figure 6. Ensemble classification precision results.

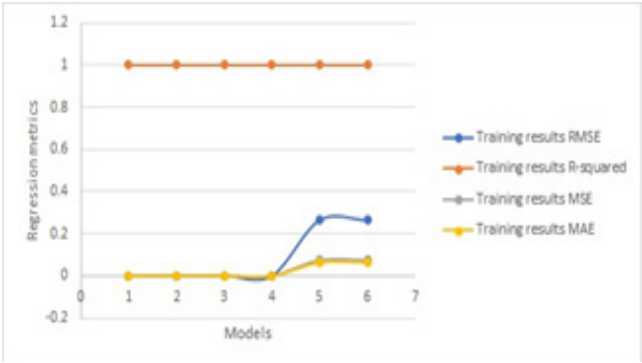

Figure 9. The outcomes of training for all regression models.

machine-learning framework. The model underwent training and testing through the utilization of repeated K-fold cross-validation.

The accuracy results for all the categorisation models used are presented in Figures 3 to 8. The decision trees had superior accuracy scores compared to the other models, attaining 97.30% and 99.50% in the training and testing outcomes, respectively. The results achieved by this model exceeded those of all other categorisation methods. On the other hand, the testing results revealed that the accuracy score of the Support Vector Machines (SVM) method was notably lower, measuring 15.30%. The suboptimal outcome can be attributed to the need for more SVM algorithms in handling massive datasets

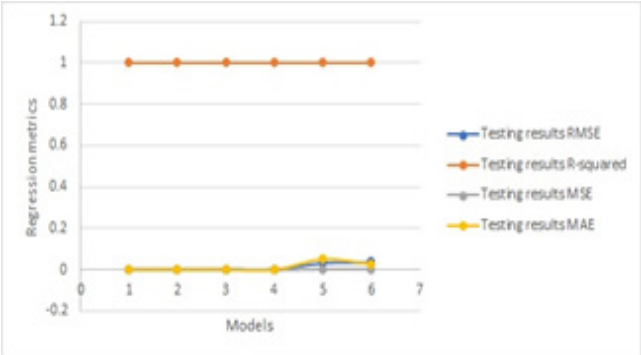

Figure 10. The outcomes of the regression models were assessed by testing.

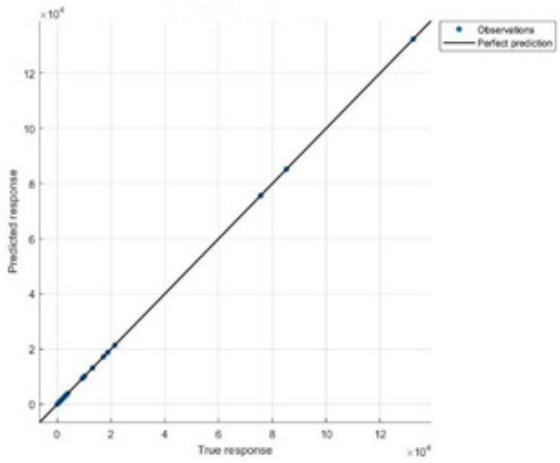

Figure 11. Plotting predicted vs. actual training model data

and its diminished efficacy in the presence of significant amounts of noise within the datasets.

However, it is worth noting that the remaining categorisation models exhibited robust performance, as seen by their respectable accuracy ratings. The Naïve Bayes algorithm demonstrated accuracy rates of 92.21% and 91.80% in the training and testing phases, respectively. The findings of the ensemble classifier indicate a performance of 94.97% and 97.73% in the training and testing phases, respectively. Similarly, the neural network attained 79.11% and 80.77% accuracy rates in the training and testing phases. Additionally, the kernel approximation approach provided accuracy rates of 81.14% and 80.90% in the training and testing phases, respectively. As

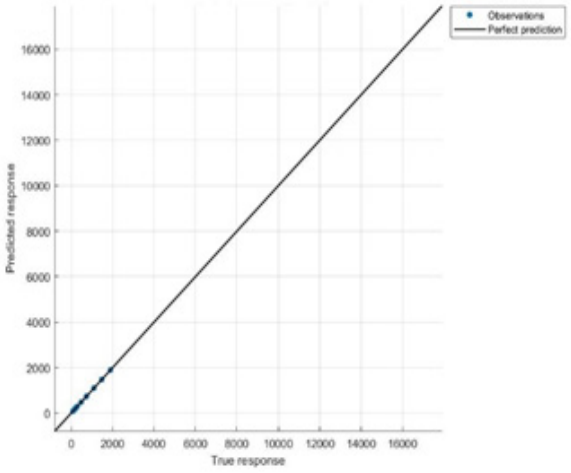

Figure 12. Plotting predicted vs. actual data for model testing.

mentioned earlier, the results underscore the enhanced efficacy of decision trees concerning other categorisation models while shedding light on the constraints inherent in the support vector machine approach. The efficacy of the remaining classification models in predicting the mine water quality index is further substantiated by their notable accuracy ratings.

Linear, Interaction, Robust, Stepwise, Rational Quadratic Gaussian Process Regression Models, and Squared Exponential Gaussian Process Regression Models are displayed in Figures 9 and 10, respectively. Root mean squared error, coefficient of determination, mean squared error and mean absolute error are all displayed graphically in the illustrations provided. Significantly, an R-squared score of 1 signifies a substantial degree of precision, indicating an impeccable forecast. The findings derived from the data mining methodology used in this investigation exhibit the most accurate projected results and contribute to the prediction of the mine water quality index.

The study built a predictive model for real-time mine water quality forecasts using regression learning. Figures 11 and 12 show this. Graphs with a linear trendline and blue data points show datasets before and after model testing. In Figures 11 and 12, the "predicted vs. actual plot" shows a flawless prediction to support the regression metrics.

4.0 Conclusion

The examination of several water quality indices has brought attention to the importance of the mine water quality index, which seeks to evaluate mine water quality using a single value thoroughly. The primary objective of this study was to assess the mathematical formulas used to calculate the mine water quality index and apply them to anticipate real-time water quality monitoring outcomes using collected datasets. Moreover, this study has presented and delineated many data mining methodologies that remain unexplored in the Indian mining industry, underscoring the necessity for additional investigation and adoption. This study showcases the utilization of the mathematically derived weighted water quality index to ascertain crucial water quality indicators inside the mining zone. As mentioned earlier, the approach has considerable promise as a beneficial instrument inside the Indian mining sector. By implementing this methodology, mining enterprises may get significant knowledge of water quality circumstances and make well-informed judgments to alleviate possible detrimental effects. This work establishes a fundamental basis for using data mining techniques to construct and predict the mine water quality index. It emphasizes the significance of using these methodologies inside the Indian mining industry to improve practices related to water quality management. Further investigation is required to continue exploring and refining these methodologies, allowing enhanced water quality evaluation and adopting sustainable mining practices within the Indian context.

5.0 References

1. Islam R, Faysal SM, Amin R, Juliana FM, Islam MJ, Alam J, Hossain MN, Asaduzzaman M. Assessment of pH and Total Dissolved Substances (TDS) in the commercially available bottled drinking water. IOSR Journal of Nursing and Health Science. 2017; 6(5):35-40.
2. Tyagi S, Sharma B, Singh P, Dobhal R. Water quality assessment in terms of water quality index. American Journal of Water Resources. 2013; 1(3):34-8. https://doi.org/10.12691/ajwr-1-3-3
3. Executive Summary of EIA/EMP: Kandri Manganese Mine. Available at http://mpcb.ecmpcb.in/notices/pdf/kandri.pdf (Accessed 14 September 2021)
4. Executive Summary of EIA/EMP: Munsar Manganese Mine. Available at http://mpcb.ecmpcb.in/notices/pdf/exe-summary-moil-nagpur.pdf (Accessed 26 September 2021)
5. United State EPA 816-F-09-004, May 2009, http://water.epa.gov/drink/contaminants/upload/mcl-2.pdf (Accessed 12 October 2021).
6. Hydrology and Water Resources Information System for India, http://117.252.14.242/rbis/india_information/water%20quality%20standards.htm (Accessed 21 October 2021).
7. Jain SK, Agarwal PK, Singh VP. Hydrology and water resources of India. Springer Science and Business Media; 2007.
8. World Health Organization. Guidelines for drinking-water quality: First addendum to the fourth edition.
9. Rajagopalan B, Lall U. A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resources Research. 1999; 35(10):3089-101. https://doi.org/10.1029/1999WR900028
10. Bessler FT, Savic DA, Walters GA. Water reservoir control with data mining. Journal of Water Resources Planning and Management. 2003; 129(1):26-34. https://doi.org/10.1061/(ASCE)0733-9496(2003)129:1(26)
11. Hyvönen S, Junninen H, Laakso L, Dal Maso M, Grönholm T, Bonn B, Keronen P, Aalto P, Hiltunen V, Pohja T, Launiainen S. A look at aerosol formation using data mining techniques. Atmospheric Chemistry and Physics. 2005; 5(12):3345-56. https://doi.org/10.5194/acp-5-3345-2005
12. Palani S, Liong SY, Tkalich P. An ANN application for water quality forecasting. Marine Pollution Bulletin. 2008; 56(9):1586-97. https://doi.org/10.1016/j.marpolbul.2008.05.021
13. Mucherino A, Papajorgji P, Pardalos PM. A survey of data mining techniques applied to agriculture. Operational Research. 2009; 9:121-40. https://doi.org/10.1007/s12351-009-0054-6
14. Gibert K, Rodríguez-Silva G, Rodríguez-Roda I. Knowledge discovery with clustering based on rules by states: A water treatment application. Environmental Modelling and Software. 2010; 25(6):712-23. https://doi.org/10.1016/j.envsoft.2009.11.004
15. Gazzaz NM, Yusoff MK, Aris AZ, Juahir H, Ramli MF. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Marine Pollution Bulletin.

2012; 64(11):2409-20. https://doi.org/10.1016/j.marpolbul.2012.08.005
16. Motamarri S, Boccelli DL. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water Research. 2012; 46(14):4508-20. https://doi.org/10.1016/j.watres.2012.05.023
17. Radojević ID, Stefanović DM, Čomić LR, Ostojić AM, Topuzović MD, Stefanović ND. Total coliforms and data mining as a tool in water quality monitoring. African Journal of Microbiology Research. 2012; 6(10):2346-56. https://doi.org/10.5897/AJMR11.1346
18. Verma A, Wei X, Kusiak A. Predicting the total suspended solids in wastewater: A data-mining approach. Engineering Applications of Artificial Intelligence. 2013; 26(4):1366-72. https://doi.org/10.1016/j.engappai.2012.08.015
19. Kovács J, Kovács S, Magyar N, Tanos P, Hatvani IG, Anda A. Classification into homogeneous groups using combined cluster and discriminant analysis. Environmental Modelling and Software. 2014; 57:52-9. https://doi.org/10.1016/j.envsoft.2014.01.010
20. Liu M, Lu J. Support vector machine- an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? Environmental Science and Pollution Research. 2014; 21:11036-53. https://doi.org/10.1007/s11356-014-3046-x
21. Mohammadpour R, Shaharuddin S, Chang CK, Zakaria NA, Ghani AA, Chan NW. Prediction of water quality index in constructed wetlands using support vector machine. Environmental Science and Pollution Research. 2015; 22:6208-19. https://doi.org/10.1007/s11356-014-3806-7
22. Babbar R, Babbar S. Predicting river water quality index using data mining techniques. Environmental Earth Sciences. 2017; 76:1-5. https://doi.org/10.1007/s12665-017-6845-9
23. Water Quality of Medium and Minor Rivers - 2019 Data as received from S.P.C.B.'s/P.C.C.'s under N.W.M.P. Available at http://www.cpcbenvis.nic.in/waterpollution/2019/Water_Quality_MediumMinor_River_2019.pdf (Accessed 13 October 2021)
24. Loh WY. Classification and regression trees. Wiley interdisciplinary reviews: Data mining and knowledge discovery. 2011; 1(1):14-23. https://doi.org/10.1002/widm.8
25. Kapil D, Tyagi P, Kumar S, Tamta VP. Cloud computing: Overview and research issues. In2017 International Conference on Green Informatics (ICGI), IEEE. 2017; 71-6. https://doi.org/10.1109/ICGI.2017.18
26. Tsai WT, Shao Q, Sun X, Elston J. Real-time service-oriented cloud computing. In 2010 6th World Congress on Services. IEEE. 2010; 473-8. https://doi.org/10.1109/SERVICES.2010.127
27. Significance of the R and D projects in MOIL: Sustainable Development Framework – Environment and Patent. 2022. Available at: https://www.moil.nic.in/userfiles/file/InvRel/Annual_Report_2020-21.pdf
28. IS I. Indian standard specification for drinking water. Google Scholar. 2012; 10500(1).
29. Ali H, Salleh MN, Saedudin R, Hussain K, Mushtaq MF. Imbalance class problems in data mining: A review. Indonesian Journal of Electrical Engineering and Computer Science. 2019; 14(3):1560-71. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563
30. Melo F. Encyclopedia of systems biology; 2013.
31. Guide Manual: Water and Waste Water, Central Pollution Control Board, New Delhi; 2021. Available at: http://www.cpcb.nic.in/upload/Latest/Latest_67_guidemanualw&wwanalysis.pdf
32. Machine Learning Crash Course. 2021. Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
33. Chauhan A, Singh S. Evaluation of Ganga water for drinking purpose by water quality index at Rishikesh, Uttarakhand, India. Report and opinion. 2010; 2(9):53-61.
34. Chowdhury RM, Muntasir SY, Hossain MM. Water quality index of water bodies along Faridpur-Barisal road in Bangladesh. Glob Eng Tech Rev. 2012; 2(3):1-8.
35. Rao CS, Rao BS, Hariharan AV, Bharathi NM. Determination of water quality index of some areas in Guntur District Andhra Pradesh; 2010.
36. Balan I, Shivakumar M, Kumar PM. An assessment of groundwater quality using water quality index in Chennai, Tamil Nadu, India. Chronicles of Young Scientists. 2012; 3(2):146. https://doi.org/10.4103/2229-5186.98688
37. Brown RM, McClelland NI, Deininger RA, O'Connor MF. A water quality index—crashing the psychological barrier. Indicators of Environmental Quality: Proceedings of a symposium held during the AAAS meeting in Philadelphia, Pennsylvania, Springer US.

1972; 173-82. https://doi.org/10.1007/978-1-4684-1698-5_15
38. MathWorks. Linear Regression. 2022. Available at: https://in.mathworks.com/help/MATLAB/data_analysis/linear-regression.html
39. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science. 2021; 7:e623. https://doi.org/10.7717/peerj-cs.623
40. Wright S. Correlation and causation; 1921.
41. Frost J. Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Minitab blog.
42. Sammut C, Webb GI. editors. Mean Absolute Error; 2010a.
43. Sammut C, Webb GI. Mean squared error. Encyclopedia of Machine Learning. 2010b; 653. https://doi.org/10.1007/978-0-387-30164-8