Regression Analysis and Correlation analysis for Prediction of Environmental Quality

C. Kaleeswari[1], K. Kuppusamy[2], A. Senthilrajan[3]

[1]Research Scholar, Dept. of Computational Logistics, Alagappa University, Tamilnadu.

[2]Formerly Professor and Head (i/c), Dept. of Computational Logistics, Alagappa University, Tamilnadu.

[3]Professor and Head, Dept. of Computational Logistics, Alagappa University, Tamilnadu.

[1]kalees94chinna@gmail.com, [2]kkdiksamy@yahoo.com, [3]agni_senthil@yahoo.com

Abstract:

Air is necessary for every living thing to breathe, be it plant, animal and human. Smoking, embers, coal ash, chemicals powder, gases, and aromas are all of wastes released into the air, either individually or in combination. Also Pollution caused by anthropogenic activities in the surface water leads to high demand for water. Machine Learning takes the lead in forecasting the quality of the air and water in environmental monitoring. For environmental protection, it is more crucial to predict gaseous pollutants in the air and Physio-chemical contaminants in the water. In this article, the environmental assessment forecasting is made using the Ordinary Least Squares model. Also Pearson Correlation Coefficient (PCC) technique is employed to find correlations between quality indicators from both datasets of air molecules and water molecules. Based on the prediction level and error rate, the results of this work portray that the Least Squares method provides good results.

Keywords: Air Quality, Water_Potabilty, Environmental Assessment, Linear Regression, Ordinary Least Squares.

DOI Number: 10.48047/nq.2022.20.19.NQ99195 **NeuroQuantology2022;20(19): 2314-2322**

Introduction

Air pollution and water pollution are two issues that are intertwined with globalization. These two issues highlighted in the research arena of environmental engineering [1–2] [20]. The World Bank's Environment Performance Index-2022 ranks India last. This means that India is among the countries with the worst environmental health. India ranks fifth out of 180 countries, with a score of 18.9. Statistics show that 63 percent of India's population is severely affected by air pollution. However, a recent study has found that women are more affected by air pollution than men. This study presented at the European Respiratory Society International Congress in Barcelona, Spain, found that breathing diesel fumes from vehicles caused changes in women's blood cells. 5 middle-aged men and women participated in this study. In examining them, it was found that both men and women were likely to suffer from diseases such as inflammatory diseases and heart diseases due to air smoke [23-25]. But apart from these, scientists have been shocked to see that many more diseases occur in women.

Environmentalists are developing their new methods by utilizing emerging methodologies to focus and predict their effects. Pollution from industrial and residential sources is a concern in the environment. Water-related pathogens and airborne diseases are caused because of the reason of pollution. An industrial waste contributes more to pollution than domestic wastewater, and the mineral industries generate a large portion of these industrial effluents [3-8].

Regression and statistical approaches are omnipresent in the real-time applications. People from various professions are attempting to use statistics to focus on making their jobs easier. Machine learning algorithms are the driving force behind such widespread use of statistics and regression analysis [17]. Regression is a technique for simulating a line of best fit using prognostic value [5] [7]. Linear Regression is a rudimental algorithm with which every Machine Learning aficionado begins.

This article will concentrate on the challenges produced by water and air defilement. The rest of this article is broken down into subsequent portions.

1. Data source: All details regarding the dataset used in this study can be found here. Molecules in the input dataset, records and its description with criteria are also given in this portion.

2. Bivariate analysis: Analysing the association betwixt the parameters in the dataset by using PCC explored.

3. Methodology: Mathematical expressions (Ordinary Least Squares Method, which was derived using Hebb's Rule) presents in this portion. Also It contains fundamental theory of regression analysis (LR Method).

4. Results and Discussion: Prediction performance and error rate provide in this portion. Finally, the conclusion portion describes LR's performance for prediction and future direction.

Applied Methodology

A. Pearson Correlation Matrix

We analysed the PCC method to assess the air and water quality variables in this part. This allowed us to demonstrate the correctness of the Machine Learning technique, which is focused on selecting features and providing justifications for selecting appropriate parameters for forecasting fresh data [22]. The correlation matrix [5] was calculated once the data sampling was estimated by utilizing the below formula.

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \qquad (1)$$

B. Preliminaries

Hebb's rule has been used to explain it. W, which is referred to as the weight matrix, can be obtained to reduce the error.

$$E_1 = \sum_{v=1}^{n} ||d_v - o_v||^2 \qquad (2)$$

where

$$d_v = \begin{matrix} d_{v1} \\ d_{v2} \\ \vdots \\ d_{vm} \end{matrix} \qquad (3)$$

is the output for craved vector;

$$o_v = \begin{pmatrix} o_{v1} \\ o_{v2} \\ \vdots \\ o_{vm} \end{pmatrix} \qquad (4)$$

is the pattern's output vector, and n is the number of patterns. To reduce the error rate of this work, the Least Squares method (Widro – Hoff law) may be used to obtain the weight matrix w.

When a specific pattern is observed

$$x_v = [x_{v1}, x_{v2}, \ldots, x_{vm}] \qquad (5)$$

When a new node is added to the network, the result yv=Wxv should be as close to the desired vector dv as possible, where v = 1, 2,..., n. To reduce the MSE, the W weight should be chosen (Mean Squared Error).

From the equation (1)

$$E_1 = \sum_{v=1}^{n} ||d_v - o_v||^2$$

$$= \sum_{v=1}^{n} \{[d_{v1} - \sum_i w_{li}\, x_{ki}]^2 + \ldots + [d_{vm} - \sum_i w_{mi}\, x_{vi}]\}$$

$$y_{v1} = w_{l1}x_{v1} + w_{l2}x_{v2} + \ldots + w_{lm}x_{vm} \qquad (6)$$

$$y_{vm} = w_{m1}x_{v1} + w_{m2}x_{v2} + \ldots + w_{mm}x_{vm} \qquad (7)$$

The w that reduce E_1 are obtained by differentiating E_1 with respect to wil for all I and l values and equating

$$\frac{\partial E_1}{\partial w_{il}} = -2\sum_{v-1}^{n}(d_{vi} - \sum_{q=1}^{m} w_{iq}\, x_{vq})x_{vl} = 0 \qquad (8)$$

which we obtain

$$\sum_{v=1}^{n} d_{vi}x_{vl} = \sum_{v=1}^{n}\sum_{q=1}^{m} w_{iq}x_{vq}x_{vl} \qquad (9)$$

$v = \{1, \dots, n\};\ q = \{1, \dots, m\};\ i = \{1, \dots, m\};$ and $l = \{1, \dots, m\};$

The equation is as follows:

$$\sum_{v=1}^{n} d_{vi}x_{vl} = \sum_{q=1}^{m} w_{iq}\left(\sum_{v=1}^{n} x_{vq}x_{vl}\right) \tag{10}$$

This derivation can be combined to form a mono matrix

$$\sum_{v=1}^{n} d_v\, x_v^T = w \sum_{v=1}^{n} x_v x_v^T \tag{11}$$

If $\sum_{v=1}^{n} x_v x_v^T$ is invertible, the w that minimizes the error is bestowed

$$w = \left(\sum_{v=1}^{n} d_v\, x_v^T\right)\left(\sum_{v=1}^{n} x_v\, x_v^T\right)^{-1} \tag{12}$$

The equation can be explained using the Hebbian rule weight matrix $\sum_{v=1}^{n} d_v\, x_v^T$ and the inverse of $\sum_{v=1}^{n} x_v\, x_v^T$.

Let X be the matrix whose columns are the input patterns, and D be the matrix whose columns are the input that agrees with the desired output vectors. Finally, we had the equation written down as

$$w = DX^T(XX^T)^{-1} \tag{13}$$

This Mathematical formulation derived from the usage of Data Mining concepts and techniques includes in Ref [21].

C. Linear Regression

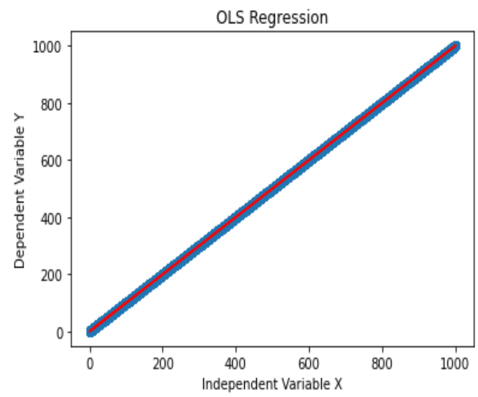
In Machine Learning, linear regression is probably the most well-known and well-understood technique. It's a linear model. The specified input (X) and the single output (Y) are assumed to communicate linearly in this paradigm (Y). Specifically, Y is derived from the input's linear connection (X). For example, in the case of a simple regression problem (a single X and a single Y), the model formation is represented as follows:

$$Y = a + bX \tag{14}$$

Where X is the regression coefficient and Y is the dependent variable. The slope of the line is b, and the intercept is a (the value of y when x = 0). A useful numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the linear association of the actual observations for the two factors. The linear regression model's formula involves assigning one factor value to each column, which is a coefficient. In addition, a coefficient is added, providing the line that moves above and below on a plot that is represented in a two-dimension. It is commonly referred to as the intercept in OLS.

D. Ordinary Least Squares Method

There are a variety of regression models available, the most popular of which is Ordinary Least Squares (OLS). Only the numeric values are acceptable for working with this technique and the outcome of this work is also numeric. When there are multiple inputs, we can use Ordinary Least Squares to estimate the coefficient values. The Ordinary Least Squares method attempts to minimize the mean squared residuals. This means that, given a regression line thru the data, we determine the distance between each data point and the regression coefficient, measure it, and add all the squares of the errors. Ordinary least squares tend to reduce this quantity.



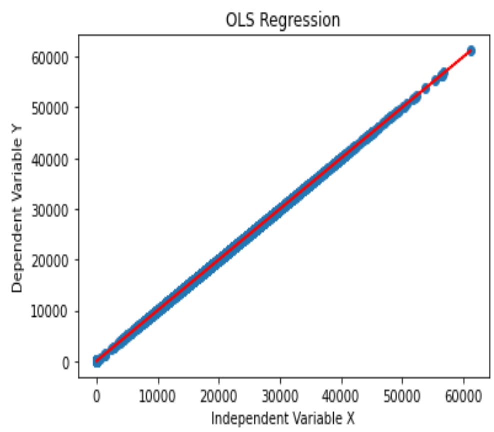


Figure 1: Best fit for Water Potability

Figure 2: Best fit for Air Quality

This system describes the data as a matrix and employs basic mathematical operations to evaluate the optimal coefficient values. It implies that almost all the data should be readily accessible, as well as sufficient memory to match the data by performing mathematical calculations. This method is extremely fast to compute. The OLS approach is used to predict water quality parameters for time series analysis in this article.

Materials

E. Data Sourcing

The Air Quality dataset gathered contains 13 attributes, including PM2.5 and PM10 (fine particulate matter), Nitrogen Monoxide or Nitric Oxide, Nitrogen dioxide, Nitrogen Oxides, Gaseous ammonia, Carbon Monoxide, Sulphur dioxide, Ground-level ozone (O3), Benzene, Toluene, Xylene, and Air Quality Index. The Water Potability dataset includes 10 parameters: pH, Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic Corban, Trihalomethanes, Turbidity, and Potabilit

Table 1: Air Quality Data Description

StationId	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
AP001	01/01/2020	59.64	88.85	1.67	12.12	7.81	14.99	0.77	17.53	57.26	0.89	3.38	0.12	96	Satisfactory
AP001	04/01/2020	22.79	38.35	2.27	18.79	11.83	14.34	0.64	19.87	23.59	0.6	5.88	0.24	47	Good
AP001	07/01/2020	69.49	97.7	1.95	12.04	8	20.46	0.68	17.89	59.37	0.71	1.24	0.1	109	Moderate
AP005	03/01/2019	148.04	285.83	25.44	103.9	75.95	16.96	1.38	29.16	93.29	5.72	9.61	4.75	319	Very Poor
AP001	22/01/2019	100.34	165.55	20.01	60.14	48.26	20.43	0.85	13.57	24.82	0.06	0.08	0.1	230	Poor

Table 2: Water Potability Data Description

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
9.44513	145.8054	13168.53	9.444471	310.5834	592.659	8.606397	77.57746	3.875165	1

0 is denoted as potable and 1 is denoted as not potable in the potability column identified from Table

Result and Discussion

A. Correlation between the Air Quality Parameters

$PM_{2.5}$ portrayed the positive relationship with PM_{10} 0.83, NO 0.38, NO_2 0.36, NO_x 0.37, NH_3 0.42, CO 0.33, SO_2 0.19, O_3 0.08, Benzene 0.33, Toluene 0.29, and Xylene 0.08. PM_{10} portrayed the positive connection with NO 0.41, NO_2 0.41, NO_x 0.41, NH_3 0.37, CO 0.41, SO_2 0.25, Benzene 0.32, Toluene 0.31, and Xylene 0.04. NO portrayed the positive association with NO_2 0.49, NO_x 0.88, NH_3 0.29, CO 0.33, SO_2 0.13, Benzene 0.31, Toluene 0.31, and Xylene 0.04. NO_2 portrayed the positive link with NO_x 0.62, NH_3 0.35, CO 0.24, SO_2 0.23, O_3 0.12, Benzene 0.31, Toluene 0.31, and Xylene 0.04. NO_x portrayed the positive correlation with NH_3 0.31, CO 0.30, SO_2 0.16, O_3 0.01, Benzene 0.29, Toluene 0.26, and Xylene 0.03.

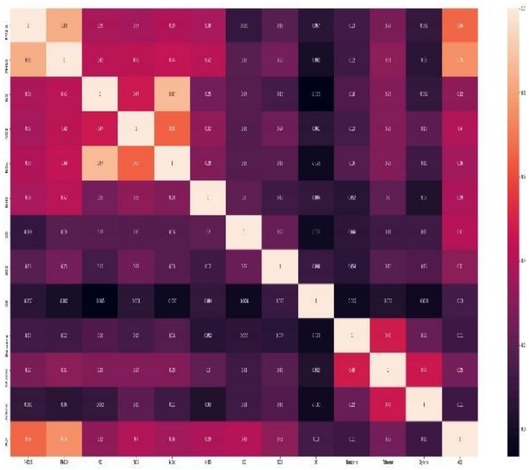


Figure 3: Correlation Map of Air Quality

NH_3 portrayed the positive association with NH_3 0.31, CO 0.22, SO_2 0.09, O_3 0.09, Benzene 0.18, Toluene 0.23, and Xylene 0.03. CO portrayed the positive connection with SO_2 0.08, Benzene 0.21, Toluene 0.21, and Xylene 0.09. SO_2 portrayed the positive relationship with O_3 0.14, Benzene 0.16, Toluene 0.13, and Xylene 0.29. O_3 portrayed the invertible correlation with PM_{10} -0.04, NO -0.04, CO -0.02, Benzene -0.01 and positive relationship with Toluene 0.05, and Xylene 0.05. Benzene portrayed the positive association with Toluene 0.45, Xylene 0.39 and Toluene portrayed the positive connection with Xylene 0.44.

B. Correlation between the Air Quality Parameters

PH portrayed the positive correlation with Trihalomethanes 0.03, Organic_Corban 0.04, Conductivity 0.02, Sulphate 0.02, Hardness 0.082 and invertible relationship with Solids -0.089, Sulphate -0.09, Chloramines -0.013. Positive association invented between Hardness and Organic_Corban is 0.0036 and negative association with Solids -0.047, Chloramines -0.03, Conductivity -0.024, Trihalomethanes -0.006. Solids portrayed the negative association with Chloramines -0.07, Trihalomethanes -0.023.

Sulphate portrayed the invertible relationship with Hardness -0.03, Solids -0.027, Conductivity -0.018 and Trihalomethanes -0.012 Solids portrayed the positive relationship with Conductivity 0.014, Organic_Corban 0.01, Turbidity 0.02. Chloramines portrayed the negative connection with Conductivity -0.02, Organic_Carbon -0.013 and positive correlation with Sulphate 0.027, Trihalomethanes 0.017, Turbidity 0.024.

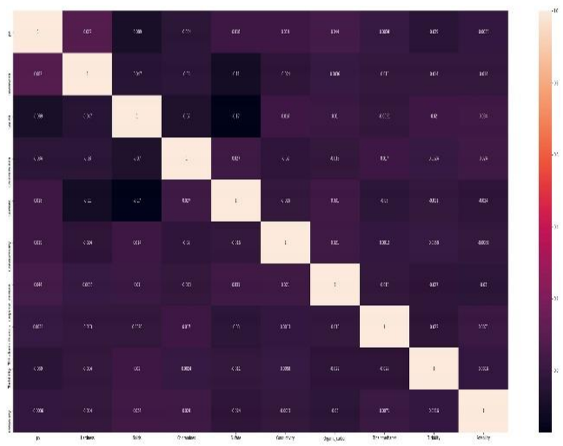


Figure 4: Correlation map of Water Potability

Conductivity portrayed the positive relationship with Organic_Corban 0.021, Trihalomethanes 0.0013, Turbidity 0.0058. Negative correlation identified between the Organic_Corbon and Trihalomethanes is -0.013, with Turbidity -0.027. Also another invertible association found betwixt the Trihalomethanes and Turbidity is -0.019.

A. Predictive Performance of OLS

Cross-referencing the predictions of water quality parameters and air quality variables is done using the results of the regression technique. Evaluation measures such as the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) are used to estimate the method's ability to produce predicted outcomes Shown in Table II. This technique was created using the open-source Anaconda Navigator (anaconda3), which is the most well-known and user-friendly environment for Python-based Machine Learning, Deep Learning, and Data Science applications [16]. The findings confirm the hypothesis that a regression technique, such as the Least Squares approach, can be used to more accurately predict environmental assessments, meaning better air and water management [9].

Table 3: Regression Results

Data		MAE	MSE	RMSE
Water	Test	0.047	0.023	0.048
	Train	0.029	0.027	0.049
Air	Test	0.068	0.087	0.091
	Train	0.068	0.087	0.093

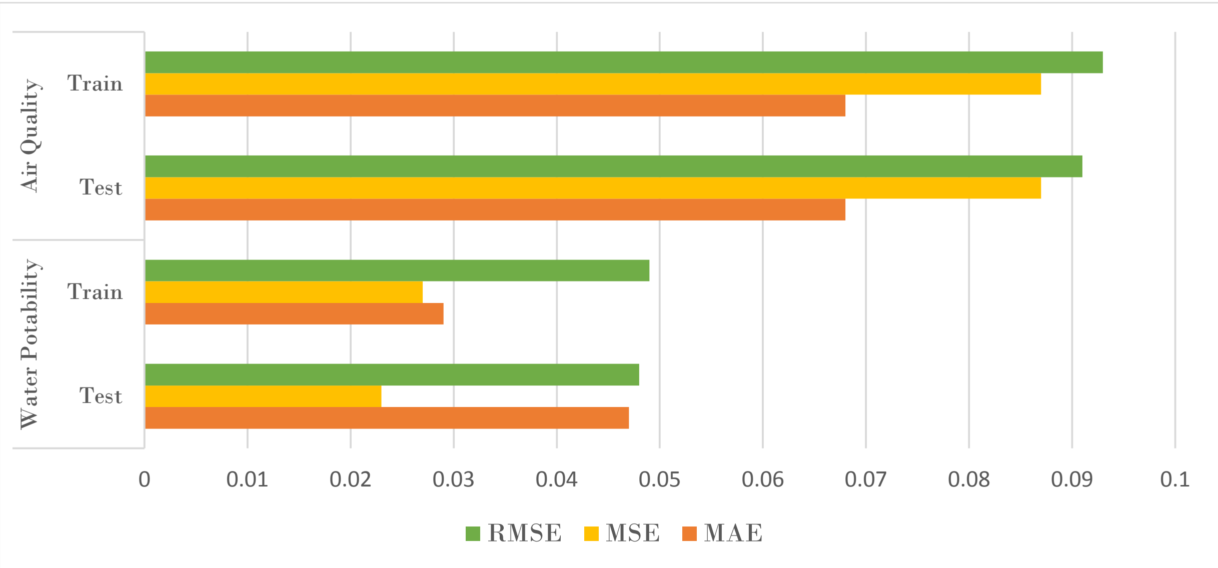


Figure 5: Performance analysis of OLS

Conclusion

Regression analysis and correlation analysis of environmental assessment were used in this paper to forecast the minimal number of parameters. Regression analysis is frequently proposed as a model; they still produce satisfying performance when integrated with computational intelligence [11]. For issues involving these processes, using the performance of linear regression may be reduced. It's a simple work and was utilized to obtain the best line. The model delivers moderate results in terms of error rate. For the reason of moderate results got from OLS method, we will try to implement by using AI (Artificial Intelligence) or Hybrid learning methods for higher prediction accuracy in the future. Another feature of that futuristic model is to examine the ability for increasing the number of prediction parameters.

Acknowledgement

This work entitled on "Regression Analysis and Correlation analysis for Prediction of Environmental Quality" done with the financial assistance from RUSA 2.0 grant sanctioned vide letter No.F.24- 51 / 2014-U, policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt 09.10.2018.

Reference:

[1] Ioannis manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos and Eugenia Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," *Frontiers in Public Health*, vol. 8, 2020.

[2] Mohammad Ail Aman, Mohd Sadiq Salman, Ali and P. Yunus, "COVID – 19 and its impact on environment: Improved pollution levels during the lockdown period – A case from Ahmedabad, India,"*Remote Sensing Applications: Society and Environment*, 2020.

[3] Kofi Owusu Ansah Amano, Eric Danso-Boateng, Ebenezer Adom, Desmond Kwame Nkansah, Ernest Sintim Amoamah and Emmanuel Appiah-Danquah, "Effect of waste landfill site on surface and ground water drinking quality," *Water and Environment Journal*, pp. 715-729, 2021.

[4] Di Wu, Hao Wang and Razak Seidu, "Smart data driven quality prediction for urban water source," *Future Generation Computer Systems*, pp. 418–432, 2020.

[5] Marlon Valentini, Gabriel Borges dos Santos and Bruno Muller Vieira, "Multiple linear regression analysis (MLR) applied for modelling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul-Brazil," *SN Applied Sciences*, 2021.

[6] NimishaWagle, Tri Dev Acharya and Dong Ha Lee, "Comprehensive Review on Application of Machine Learning Algorithms for Water Quality Parameter Estimation Using Remote Sensing Data," *Sensor and Materials*, vol. 32, No. 11, pp. 3879-3892, 2020.

[7] Arun Kumar Shrestha and Nabin Basnet, "The Correlation and Regression Analysis of Physicochemical Parameters of River Water for the Evaluation of Percentage Contribution to Electrical Conductivity,"*Journal of Chemistry*, vol. 2018.

[8] Mate Krisztian Kardon and Adrienne Clement, "Predicting small water courses' physico-chemical status from watershed characteristics with two multivariate statistical methods," *Open Geosci*, vol. 12, pp. 71-84, 2020.

[9] Dostdar Hussain and Aftab Ahmed Khan, "Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan," *Earth Science Informatics*, vol. 13, pp. 939-949, 2020.

[10] S. I. Abba, Sinan Jasim Hadi, Saad Sh. Sammen, Sinan Q. Salih, R.A. Abdulkadir, Quoc Bao Pham and Zaher Mundher Yaseen, "Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination," *Journal of Hydrology*, 2020.

[11] Xin Zhang, Yuqi Liu and Lin Zhou, "Correlation Analysis between Landscape Metrics and Water Quality under Multiple Scales," *International Journal of Environmental Research and Public Health*, 2018.

[12] Haixia Lin, Jianhong Cui and Xiangwei Bai, "Feature Extraction of Marine Water Pollution Based on Data Mining," *Symmetry*, 2021.

[13] Parveen Sihag, Munish Kumar and Balraj Singh, "Assessment of infiltration models

developed using soft computing techniques," *Geology, Ecology, and Landscapes*, 2020.

[14] Sunsook jang, Hyunseoji, KyoSuh and Hakkwankim, "Investigation of correlation between surface runoff rate and stream water quality," *Water Supply*, 2021.

[15] Kangyang chen, Hexia Chen, Chuanlong Zhou, Yichao Huang, Xiangyang Qi, RuqinShen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng and Hongqiang Ren, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," *Water Research*, 2020.

[16] Ali El Bilali and Abdeslam Taleb, "Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment," *Journal of the Saudi Society of Agricultural Sciences*, vol.19, pp. 439-451, 2020.

[17] Faridah Othman, M. E. Alaaeldin , Mohammed Seyam, Ali Najah Ahmed, Fang TennTeo, Chow Ming Fai, Haitham Abdulmohsin Afan, Mohsen Sherif, Ahmed Sefelnasr and Ahmed El-Shafie, "Efficient river water quality index prediction considering minimal number of inputs variables," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, No. 1, pp. 751-763, 2020.

[18] Umair Ahmed, Raffia Mumtaz, Hira Anwar, Assad A. Shah, RabiaIrfan and Jose Garcia-Nieto, "Efficient Water Quality Prediction Using Supervised Machine Learning," *Water*, 2019.

[19] Tord Kjellstrom, Madhumita Lodh, Tony McMichael, Geetha Ranmuthugala, Rupendra Shrestha, and Sally Kingsland, "Air and Water pollution: Burden and Strategies for Control," *Disease Control Priorities in Developing Countries,* Chapter 43.

[20] Louis C. McCabe, M. Allen Pond and E. Neil Helmers, "Interrelationship of Air Pollution and Water Pollution," *JSTOR Sewage and Industrial Wastes*, vol. 24, No. 1, pp. 83-91, 1952.

[21] Jiawei Han, Micheline Kamber and Jianpei, "Data Mining Concepts and Techniques," *Elsevier*, Third Edition.

[22] Yifan Zhang, Peter Fitch, Maria P. Vilas and Peter J. Thorburn, "Applying Multi-Layer Artificial Neural Network and Mutual Information to the Prediction of Trends on Dissolved Oxygen," *frontiers in Environmental Science*, vol. 7, 2019.

[23] Mahadevappa Hemshekhar, Hadeesha Piyadasa, Dina Mostafa, Leola N. Y. Chow, Andrew J. Halayko, Neeloffer Mookherjee, "Cathelicidin and Calprotectin Are Disparately Altered in Murine Models of Inflammatory Arthritis and Airway Inflammation," *frontiers in Immunology*, vol.11, 2020.

[24] Neeloffer Mookherjee, Min Hyung Ryu, Mahadevappa Hemshekhar, Juma Orach, Victor Spicer, Christopher Carlsten, "Defining the effects of traffic-related air pollution on the human plasma proteome using an aptamer proteomic array: A dose-dependent increase in atherosclerosis-related proteins," *Environmental Research,* Vol. 209, 2022.

[25] Zorana J. Andersen, Jiawei Zhang, Jeanette T. Jørgensen, Evangelia Samoli, Shuo Liu, Jie Chen, Maciej Strak, Kathrin Wolf, Gudrun Weinmayr, Sophia Rodopolou, Elizabeth Remfry, Kees de Hoogh, Tom Bellander, Jørgen Brandt, Hans Concin, Emanuel Zitt, Daniela Fecht, Francesco Forastiere, John Gulliver, Barbara Hoffmann, Ulla A. Hvidtfeldt, W.M. Monique Verschuren, Karl-Heinz Jöckel, Rina So, Tom Cole-Hunter, Amar J. Mehta, Laust H. Mortensen, Matthias Ketzel, Anton Lager, Karin Leander, Petter Ljungman, Gianluca Severi, Marie-Christine Boutron-Ruault, Patrik K.E. Magnusson, Gabriele Nagel, Göran Pershagen, Annette Peters, Debora Rizzuto, Yvonne T. van der Schouw, Sara Schramm, Massimo Stafoggia, Klea Katsouyanni, Bert Brunekreef, Gerard Hoek, Youn-Hee Lim, "Long-term exposure to air pollution and mortality from dementia, psychiatric disorders, and suicide in a large pooled European cohort: ELAPSE study," *Environment International*, Vol. 170, 2022.