International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Deep Insight into Urban Air Quality Utilizing Neural Networks for Enhanced Prediction in Korean Cities Where Factories and Ecosystem Environments Coexists

Hyun Sim[1], Hyunwook Kim[2]
[1]Profrssor, Dep. Smart Agriculuture
Sunchon Nationa University
Suncheon, South Korea
e-mail: simhyun@scnu.ac.kr
[2]Kornerstone. Co, Ltd.
Yeosu, South Korea
e-mail: gratia@kornerstone.kr

Abstract— Increased attention is being given to air pollution in recent times. This study investigated and analyzed particulate matter data from Yeosu, Gwangyang, and Suncheon in Jeollanam-do, with a particular focus on PM2.5. Descriptive statistics, box-and-whisker plots, correlation matrices, time variations, and trend analyses were performed for this purpose. Additionally, a prediction model for PM2.5 concentrations was developed using machine learning techniques, through which future changes in air quality were forecasted.

Calculations were performed using R-based programs and R packages. Hourly PM2.5 data were obtained from air quality monitoring sites in Yeosu, Gwangyang, and Suncheon. After data preprocessing, the optimal prediction model was constructed using Random Forest and Gradient Boosting Machine from various machine learning algorithms.

The research results showed that there was more PM2.5 pollution in Gwangyang compared to Yeosu and Suncheon. The PM2.5 concentrations varied significantly across each monitoring site. Among the monitoring sites, the Yeosu site showed a higher correlation in PM2.5 with each other than other sites. Late winter and early spring showed higher PM2.5 concentrations, while summer and autumn showed lower concentrations. Weekly PM2.5 concentration fluctuations were not significantly different. Daily fluctuations showed an increase in PM2.5 concentrations during times of traffic congestion and a decrease in the afternoon. During the research period, the trend of PM2.5 concentration was generally decreasing.

The accuracy of the prediction model through machine learning was over 90%, and it is expected to assist in establishing effective response strategies for future changes in air quality. This study provided an updated and useful evaluation of recent PM2.5 air quality in Yeosu, Gwangyang, and Suncheon in Korea.

Keywords-Air Pollution, Machine Learning, Particulate Matter, Prediction Model, Statistical Analysis, Statistical Software R.

I. INTRODUCTION *(HEADING 1)*

With the alterations in global environment, the issue of air pollution has emerged as a major environmental concern worldwide. Among various environmental pollutants, particulate matter is progressively being emphasized due to its potential to penetrate the human body owing to its minute size and its linkage to various health issues. While particulate matter can originate naturally, in nations undergoing industrialization and urbanization, human activities are often pinpointed as a primary source of particulate matter.

Air pollution has established itself as a persistently highlighted environmental issue in numerous regions across the globe, including Korea. Accordingly, the rise in air pollution due to industrial development, urbanization, energy consumption, and crop burning in developing countries is receiving significant attention. Principal components of particulate matter include organic chemicals, acids, soil, and dust particles, the effects of which on the human body have already been demonstrated through various studies. Particularly, PM2.5 and PM10 are known to be closely related to health issues.

Air pollution refers to the contamination of internal or external air and arises due to any chemical, physical, or biological factor that alters the natural characteristics of the atmosphere. Common sources of air pollution include domestic combustion devices, vehicles, industrial activities, and wildfires. Even at extremely low concentrations, particulate matter pollution can negatively impact health. In fact, no threshold has been found that does not harm health. Particles less than 10 millimeters in size (PM10) can enter and reside in the respiratory

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

tract, but particles smaller than 2.5 microns (PM2.5) are more harmful to health.

Machine learning is an innovative technology for predicting and analyzing air pollution[1]. Over recent years, various machine learning strategies have been proposed to predict various air pollutants using various combinations of predictor variables. Regression analysis can assist in analyzing data and making informed[2]. When using regression models for prediction, the goal is to get predictions that are, on average, close to the actual values.

The widespread awareness of these issues has driven a pursuit of deeper understanding of particulate matter through various research endeavors. This study focuses on the cities of Suncheon, Yeosu, and Gwangyang in Jeollanam-do, Korea, which are representative tourist and ecological cities and possess a unique urban structure where factories and ecological environments coexist. In regions like Gwangyang Bay Area, there is high interest in the influence cities might have on each other when factories and ecological environments coexist. For this reason, through analyzing the particulate matter concentration in the areas of Yeosu, Gwangyang, and Suncheon in Jeollanam-do, Korea, this study aims to clarify the local air quality changes and their causes. Furthermore, utilizing the latest technology of machine learning, a model predicting future particulate matter concentrations was constructed. The results obtained through this study are expected to significantly assist local communities and policymakers in providing strategies for future air quality management.

II. RELATED WORK

Various machine learning algorithms have been proposed over recent years to address the problem of predicting PM2.5 particulate matter. In this section, research in this field will be introduced and analyzed.

A. Approach to PM2.5 Prediction Using Regression Models

Southeast Asia is a priority region for environmental pollution and haze conditions. The validation of RF spatial models worked slightly better than SVR, with statistical indicators in urban/industrial areas showing R2 = 0.76, RMSE = 11.47 g, and in suburban/rural areas R2 = 0.64, RMSE = 10.76. The aim of this study was to predict PM2.5 concentrations in Malaysia using machine learning (ML) models derived from satellite AOD (Aerosol Optical Depth) data, ground harmful emissions, and meteorological variables[3]. Calibration of SVR performed slightly better for the overall model than RF calibration, showing R2 = 0.69 and RMSE = 10.62 against observed PM2.5 concentrations. The authors suggested including gaseous pollutants from satellite remote sensing data in ML techniques to estimate PM2.5 concentrations in future research.

In the past few years, weather and traffic variables, fossil fuel usage, and industrial characteristics have all played significant roles in air pollution. The aim of this research was to examine the relationships between these variables using regression analysis methods and to predict Carbon Monoxide (CO) based on other data. The results revealed that Lasso Regression demonstrated the optimal model[4].

For years, excessive concentrations of particulate matter of sizes PM10 and PM2.5 have resulted in serious health issues. The information used to train the models was obtained from Taiwan's Air Quality Monitoring Network (TAQMN). The final results identified Gradient Boosting Regression as the optimal model. The authors employed various regression models, including linear regression, Lasso regression, Ridge regression, random forest regression, Gradient Boosting regression, and MLP regression[5].

Generally, rapid growth, urbanization, and improved living standards have greatly amplified urban air pollution. Data sets are obtained from the locations of the Central Pollution Control Board (CPCB) and the "Atmosphere and Noise Pollution Monitoring System" for research. As three bases were evaluated, the R-squared value for Gradient Boosting Regression was 0.69647 in R.K Puram, performing better than the other models. However, scientists advised that future research includes other meteorological factors, such as precipitation, minimum and maximum temperature, solar radiation, and vapor pressure, to enhance precision.

Malaysia experiences transboundary haze events annually. When solid materials suspended in the atmosphere, especially PM10, are involved, they affect humans and the environment. The study concludes with a focus on research findings that assist responsible parties in providing early warning information, and help mitigate and prevent activities to improve air quality and enhance human health during haze episodes. For deeper investigation, the dataset was received from the Malaysian Department of Environment (DOE) and the Ministry of Natural Resources and Environment. The results demonstrate that multiple linear regression showed better performance than other regression models[6].

With the increasing frequency of hazy weather, predicting the concentration of PM2.5, the major pollutant during hazy weather, has gradually become a hot topic. The overall goal of the paper is to predict PM2.5 concentration using multivariate linear regression models. For the research, the authors used data from China's air quality online monitoring and analysis platform. The results show R-Squared = 0.8782 and F-Test = 98.4152, indicating that the model fits well and can be predicted by the model[7].

Air pollution has been revealed to be an important predictive factor for predicting human health. The objective of this project

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

is to predict pollution using four complex regression algorithms and to perform a comparative analysis to determine which model is most suitable for reliably predicting air quality. Random Forest Regression succeeded in identifying peak values and, in fact, took less time to analyze than other models. For various datasets, the MAE ranged between 6% and 18%, and RMSE between 0.05 and 0.18. Random Forest Regression worked well after hyperparameter tuning[8].

The purpose of the present research is to understand the studies of other experts on air pollution and the limitations of their work. The main research gap found is that only a few studies have been reviewed for predicting PM2.5 using machine learning. Although numerous studies on PM2.5 have been conducted in various aspects and regions, PM2.5 has not yet been introduced into the field of data science. However, the development of machine-learning algorithms is widespread, and more accurate predictions can be counteracted within the algorithm to obtain a stronger response. Therefore, since it has been proven in some studies that these machine-learning models are superior to others, it was decided to implement Support Vector Regression, Decision Tree Regression, and Multiple Linear Regression in this study.

III. RESEARCH METHOD

In this section, we explore how regression models, such as Support Vector Regression, Multiple Linear Regression, and Decision Tree Regression, are implemented to determine the optimal prediction model for particulate matter PM2.5 in the air quality of Yeosu, Suncheon, and Gwangyang. The procedure begins with obtaining data and continues through data cleaning, feature selection, regression modeling, and ultimately, performance measurement.

A. Dataset

In the initial stage of this study, data regarding the air quality in Yeosu, Suncheon, and Gwangyang was collected from Kaggle and AirKorea. This data encompasses various variables, such as the concentration of several chemical substances, over the period from 2019 to 2023. Each dataset is provided as a zip file of merged csv files, from which each zip file was extracted and combined via programming to form one large dataset. The dataset consists of 467,568 rows and 12 columns. The dataset used includes region name, measurement station code, measurement station name, measurement time, SO2, CO, O3, NO2, PM10, PM25, and address values. Data mining and analysis techniques were employed to transform the collected raw data into data that can be analyzed. The data was cleaned through processes like deleting rows with null values, changing data types to integers, resolving data inconsistencies, and splitting the date into new columns to add new variables that include the year, month, day, and quarter.

B. Data Cleaning

Handling missing values was a major concern during the data cleaning process. In this study, missing values for each chemical substance concentration were replaced with the mean value of the respective variable. Additionally, the variables used for analysis were standardized to enhance the performance of the machine learning model.

C. Data Splitting

The data was split into training and validation data. In this study, the 'measurement time' variable was used as a basis for data splitting, with data from 2019 to 2021 being used as training data and, hypothetically, if data from 2022 is added, it would be used as validation data.

D. Feature Engineering

Feature selection is a method that automatically selects the features in data that contribute most significantly to the predictive variable or output of interest[9]. In the feature selection process, variables like 'SO2', 'CO', 'O3', 'NO2', and 'PM25' were chosen as independent variables, and 'PM10' was chosen as the dependent variable. This selection aligns with the research objective, which is focused on PM2.5 prediction, and was used to structure the input and output of the model. Various analytical models, especially linear approaches like linear and logistic regression, can be impacted by unnecessary features in the data. In this study, heatmap correlation was used to choose how the independent variables are correlated with the dependent variable. The heatmap provides exact correlation relationships between 0 and 1, making it understandable and easy to read. Figure 1 shows the heatmap results for the dataset.

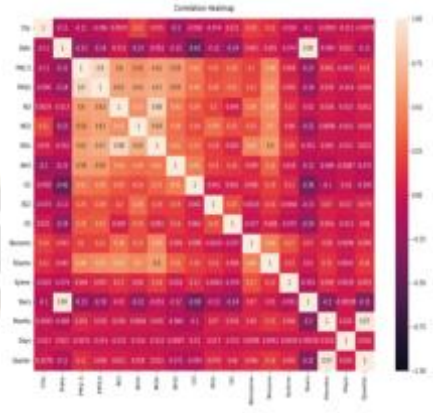

Figure 1. Correlations, heatmaps for datasets

E. Support Vector Regression

Support Vector Regression (SVR) is a supervised learning method for predicting discrete values. With a few minor modifications, SVR uses the same principles as the Support Vector Machine (SVM) for classification. In regression, the SVM is provided with a permissible error (epsilon) as an approximate estimate that has already been requested[10].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

F. Multiple Linear Regression

Multiple Linear Regression is often recognized as a quantitative approach to predict the outcome of a response variable using multiple explanatory variables. Both linear and nonlinear regression use graphs employing two or more variables to track a specific response. However, implementing nonlinear regression can be challenging since it typically relies on assumptions based on trial and error[11].

G. Decision Tree Regression

The decision tree algorithm belongs to the family of supervised learning algorithms. Unlike a few other supervised learning methods, decision tree methods can be utilized to solve both regression and classification problems. The purpose of using decision trees is to build a training model that predicts the type or quantity of the target attribute by learning the basic decision rules from past data (training data). In decision trees, the prediction of the target class for a record starts from the root of the tree, and the characteristics of the connected network are computed, with the result of the record's attribute being calculated[12].

H. Model Performance Metrics

Model evaluation is a crucial step in the model creation process. This step helps determine the optimal model that represents the data and assists in how well the chosen model will perform in the future.

I. Mean Absolute Error (MAE)

MAE represents the rate of variation between matching observations that reflect the same phenomenon. Equation 1 shows the formula for MAE[13].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (1)$$

In Equation 1, y_i represents the prediction, $x_i$represents the actual value, and n denotes the total quantity of data.

J. Coefficient of Determination (R-Squared)

R-squared is a measure of the fit of the linear regression model. It shows how much variation is explained by the external variables. R-squared reflects the strength of the association between the model and the predictor variables, offering a convenient 0-100% scale. Equation 2 demonstrates the formula for R-Squared.

$$R^2 = 1 - \frac{RSS}{TSS} \quad (2)$$

Here, 2R2 is the coefficient of determination, RSS (Sum of squares of residual) is the sum of squares of the residual, and TSS (Total sum of squares) is the total sum of squares.

K. Root Mean Squared Error (RMSE)

RMSE is a common method for quantifying the quality of fit in statistical modeling, especially in regression analysis. Equation 3 demonstrates the formula for RMSE.

$$RSME = \sqrt{\frac{\sum_{i=1}^{N}(x_i - a_i)^2}{N}} \quad (3)$$

Where N is the number of non-missing data, x_i is the actual observed time series, and a_i is the estimated time series.

L. Adjusted R-Squared

The adjusted R-squared is a version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term enhances the model more than would be expected by chance. Traditionally, adjusted R-squared is positive instead of negative and is always smaller than R-squared. Equation 4 shows the formula for implementation.

$$Adj\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \quad (4)$$

Where R^2 is the sample coefficient of determination, N represents the size of the entire sample, and p is the number of independent variables.

IV. RESULTS AND DISCUSSION

Strategies for determining the optimal model for predicting PM2.5 are discussed in this section, accompanied by performance measurements. The flow of the section begins with the performance of Support Vector Regression, moving through Multiple Linear Regression, Decision Tree Regression, Adjusted R-Squared, and overall performance. To obtain better performance metric values, all regression models typically underwent hyperparameter tuning using the grid search method.

A. Performance of Support Vector Regression (SVR)

In this experiment, the SVR regressor was utilized to learn and predict PM2.5 values. According to SVR, the error levels of MAE and RMSE are 15.36 and 33.42, respectively. This issue is likely to be mostly resolved after adjusting the hyperparameters. The MAE and RMSE values in the test set decreased to 13.10 and 23.94, respectively. Additionally, the R-Squared value slightly increased from 0.65 to 0.82 after hyperparameter tuning. Table 1 displays the results of the performance metrics.

TABLE I. SUPPORT VECTOR REGRESSION PERFORMANCE METRICS

Before parameter tuning			**After parameter tuning**		
MAE	RMSE	R^2	MAE	RMSE	R^2
15.36	33.42	0.65	13.10	23.94	0.82

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

B. Performance of Multiple Linear Regression (MLR)

In addition to SVR, MLR regression was used to predict PM2.5 levels. The results show that the MAE and RMSE of the test set are 13.29 and 23.42, respectively. After adjusting a few parameters, the MAE and RMSE of the test set slightly decreased to 13.28 and 23.42, respectively. Moreover, the R-Squared result appeared higher at 0.83 after hyperparameter tuning, from 0.82. The statistics for performance measurement are presented in Table 2.

TABLE II. MULTIPLE LINEAR REGRESSION PERFORMANCE METRICS

Before parameter tuning			After parameter tuning		
MAE	RMSE	R^2	MAE	RMSE	R^2
13.3	23.4	0.8	13.2	23.42	0.8

C. Performance of Decision Tree Regression

The final model was implemented as a decision tree regression. Utilizing default values, the decision tree was employed to predict the values of PM2.5. The MAE and RMSE results for the test set are 11.55 and 22.24, respectively. After adjusting some hyperparameters, the MAE and RMSE in the test set decreased to 10.80 and 20.70, respectively. Additionally, the adjusted R-Squared was calculated using variable interactions. The R-Squared result showed 0.84 before hyperparameter tuning and 0.86 after. The results of the performance measurement are shown in Table 3.

TABLE III. DECISION TREE REGRESSION PERFORMANCE METRICS

Before parameter tuning			After parameter tuning		
MAE	RMSE	R^2	MAE	RMSE	R^2
11.55	22.24	0.84	10.80	20.70	0.86

D. Adjusted R-Squared

TABLE IV. TABLE TYPE STYLES

Model	R^2	Adjusted R^2		
		4	8	13
SVR	0.8217	0.8214	0.8211	0.8208
MLR	0.8294	0.8291	0.8288	0.8285
DTR	0.8666	0.8620	0.8618	0.8615

Table 4 compares the R-Squared and adjusted R-Squared of the three models. It indicates that the decision tree's value is the best to choose since it has the highest value. Despite having the maximum value, R-Squared could lead to overfitting due to the chosen predictive variables, 13 in this case. Adjusted R-Squared is used to avoid overfitting, working by selecting only a limited number of independent variables. In this instance, a maximum of 13 independent variables can be investigated using the adjusted R-Squared. Therefore, as an overall study, four variables will be selected since the values for 8 and 13 significantly decrease, indicating that the added variables might not be correlated with the target variable.

TABLE V. OVERALL PERFORMANCE METRICS OF MODELS

Model	MAE	RMSE	R^2
SVR	13.1046	23.9415	0.8217
MLR	13.2893	23.4213	0.8294
DTR	10.8070	20.7084	0.8666

E. ML Model Construction: Neural Networks

A neural network was constructed as the machine learning model. This model includes multiple hidden layers and neurons, employing the 'relu' activation function. The final output layer utilized a linear activation function suitable for regression problems. The model was compiled using an optimizer and a loss function.

Figure 2. Model output classification and impact factor

F. Model Training and Performance Evaluation

As seen in Figure 3, the model was trained for 50 epochs, with a batch size set to 32. During the training process, 20% of the data was used as validation data, and the loss and Mean Absolute Error (MAE) during the learning process were visualized for evaluating the model's performance. This helps to prevent the model from being overly applied to the learning data and enhances generalization capability.

Figure 4 shows the result monitored using the validation dataset during the model training. The model's performance was continuously monitored and managed overfitting.

Loss and MAE during the training process were recorded per Epoch and visualized to evaluate the model's learning progression. This allowed for observing the learning curve and identifying the point where overfitting occurs.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

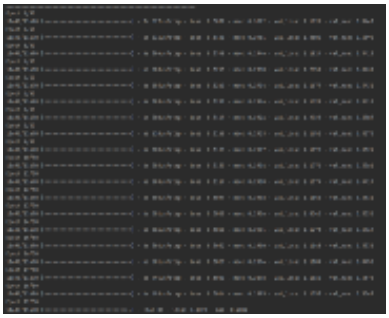

Figure 3. Training data and MAE

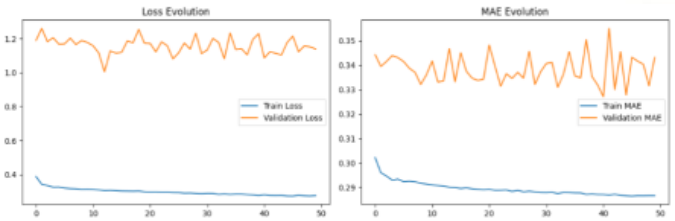

Figure 4. Loss and MAE Evaluation

V. CONCLUSION

This study conducted data analysis on the air quality of Yeosu, Suncheon, and Gwangyang, particularly exploring the optimal prediction model for PM2.5, or fine dust. Various regression models, such as Support Vector Regression, Multiple Linear Regression, and Decision Tree Regression, were implemented and compared for this purpose. The specific procedure includes data collection, data cleaning, feature selection, regression modeling, and performance measurement.

The provided dataset, which includes detailed air quality information, was collected from Kaggle and Air Korea. Through the data mining and analysis process, this study initially processed the data and transformed it into an analyzable form. Subsequently, feature selection was used to select the characteristics that contribute most significantly to the predictive variables.

Predictions for PM2.5 were performed through three regression models: Support Vector Regression, Multiple Linear Regression, and Decision Tree Regression. These models predicted the concentration of PM2.5 according to their respective methodologies, and the primary objective of this study was to compare the performance of these models and find the optimal one among them. Effective model learning requires appropriately mixing and applying the aforementioned strategies. By using various metrics and validation strategies to measure the model's performance, the model should perform well in actual environments, and consistent monitoring and improvement of the model's predictive performance are deemed necessary, especially in environmental issues like air quality prediction where accurate predictions are crucial.

The performance of the model was measured using several metrics, including Mean Absolute Error (MAE) and R-Squared. These metrics were used to evaluate the predictive performance and fitness of the model to the data.

The results of this study, especially in predicting PM2.5, reveal the usefulness and limitations of certain models, providing critical insights into the development of future air quality prediction and management strategies. Such analyses and predictions will assist policymakers and environmental scientists in understanding the causes and effects of air pollution more accurately and in devising effective strategies and plans to improve the air quality of communities. However, additional validation is required for predicting air quality in actual environments based on the results of this study, which can be conducted in future research. The research can also be expanded to enhance the model's generalization capability by including other environmental variables and data from different regions.

This study also emphasizes the importance of machine learning models and statistical analysis, which can be utilized as powerful tools in the analysis and prediction of complex environmental data. These approaches will be further utilized in the analysis and prediction of various environmental data in the future, thereby reinforcing the scientific foundation for the protection and management of our environment.

ACKNOWLEDGMENT

" This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2023-2020-0-01489) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)"

REFERENCES

[1] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, M. N. Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities, IEEE Access, vol. 7, pp. 128325-128338, September 2019. doi:10.1109/ACCESS.2019.2925082.

[2] L. Teeboom, "The Advantages of Regression Analysis & Forecasting," 8 March 2019.

[3] N. A. Zaman, K. D. Kanniah, D. G. Kaskaoutis, M. T. Latif, "Evaluation of Machine LearningModels for Estimating PM2.5 Concentrations across Malaysia," Applied Sciences, vol 11, no 16, 2021, pp.7326. dio:10.3390/app11167326.

[4] S. Abdullah, N. N. Napi, A. N. Ahmed, W. N. Mansor ,A. A. Mansor, M. Ismail, A. M. Abdullah, Z. T. Ramly, "Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia," Atmosphere, 2020, p. 289.

[5] K. S. Harishkumar, K. M. Yogesh, I. Gad, "Forecasting air pollution particulate matter (PM2.5) usingmachine learning regression models," Procedia Computer Science, 2020, pp. 2057–2066.

[6] S. Abdullah,Napi NN, Ahmed AN, Mansor WN,Mansor AA, Ismail M, AbdullahAM, Ramly ZT. Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia. Atmosphere, 2020, pp.289.

[7] J. Chen, J. Wang, "Prediction of PM2.5 concentration based on multiple linear regression," In2019 International Conference on Smart Grid and Electrical Automation (ICSGEA), 2019, pp. 457–460, IEEE.

[8] C. Srivastava, S. Singh, A. P. Singh, "Estimation of air pollution in Delhi using machine learning techniques," In 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 304–309. IEEE.

[9] R. Shaikh, "Feature Selection Techniques in Machine Learning with Python," 2018.

[10] A. Sethi , "Support Vector Regression Tutorial forMachine Learning," 2020.

[11] R. Bevans, "Multiple Linear Regression |AQuickGuide (Examples)," Scribbr, 2020.

[12] N. S. Chauhan, "Decision TreeAlgorithm, Explained,"KDnugget, 2022.

[13] S. Hiregoudar, "Ways to Evaluate Regression Models," 2020.