Machine Learning Regression Models to Predict Particulate Matter ($PM_{2.5}$)

Koogan A. L. Letchumanan and Naveen Palanichamy(✉)

Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
p.naveen@mmu.edu.my

Abstract. An increase in the quantity of fine particulates ($PM_{2.5}$) in the air is a risk to the nation's people since it can create uncontrolled repercussions such as the aggravation of cardiovascular disease and asthma. The issue of air pollution has lately surfaced as a critical concern in smart cities. The systematic technique of estimate particulate matter 2.5 using Machine Learning (ML) has received a lot of attention over the years. The main motive of the research is to employ machine learning models to find the best predicting model to forecast particulate matter $PM_{2.5}$ in air quality in smart urban. Support Vector Regression, Decision Tree and Multiple Linear regression are chosen to study the application of machine learning in this research. The outcome of the prediction from respective machine learning then will be evaluated by the performance metrics to measure performance of the models. The outcome demonstrates that Decision Tree Regression is the best fit model for our present study.

Keywords: Air pollution · Smart cities · $PM_{2.5}$ · Decision tree · SVR · Multiple linear regression

1 Introduction

A smart city is an urban municipality that uses information and communication technologies (ICT) to provide better health, transportation, and energy-related utilities to its citizens, help communities to make efficient use of its significant assets for the welfare of people [1]. Many smart cities are experiencing air pollution issues. In the modern world, urban populations have grown rapidly as a result of industrialisation and migration from rural to urban areas. The city's population expansion has resulted in a greater number of users of transportation and energy, which contributes to the city's growth of industry and autos. As a result, multiple studies have found that smog in green infrastructure is a significant impediment to the town's capacity to provide inhabitants with a better and healthier way of life.

Air pollution is defined as contamination of the interior or external atmosphere by any chemical, physical, or biological factor that alters the inherent properties of the atmosphere. Common causes of air pollution include domestic burning devices, motor vehicles, industrial operations, and forest fires [2]. Even at extremely low concentrations, small particle pollution has a negative impact on health; in fact, no threshold has been

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 458–468, 2022.
https://doi.org/10.2991/978-94-6463-094-7_36

discovered beyond which no harm to health is seen. Nanoparticles having dimensions of 10 millimetres fewer (PM_{10}) may enter and lodge respiratory system, but particles with a diameter of 2.5 microns or less ($PM_{2.5}$) are more detrimental to one's health. This enables policymakers to forecast the gains in population health that may be expected if particle air pollution is decreased.

Machine learning is a novel and innovative technology for predicting and analyzing air pollution [3]. Many machine learning strategies have been proposed in recent years for dealing with the air quality forecast issue in astute urban communities. Several ML methods have been used in recent years to predict a variety of air contaminants using various combinations of predictor factors.

Regression analysis may help you crunch the data and to make smart decisions for your organization now and in the future [4]. When using regression to create predictions is to get forecasts that are both right on average and near to the true values. To put it another way, we need unbiased and exact predictions.

The paper is starting with the related work in segment 2 in predicting $PM_{2.5}$. Then, followed with segment 3 proposes a method for explaining machine learning algorithms in greater depth, as well as the fundamental data mining pipeline. Move on to segment 4, findings, where to highlight how the study was carried out and how we used the data with the performance metrics results, and for segment 5, conclusion.

2 Related Work in Predicting $PM_{2.5}$

Many machine learning algorithms have been presented in recent years to solve the particulate matter 2.5 prediction problem in smart cities. This segment presents and analyses the related work in this field.

2.1 Approaches in Predicting PMatter2.5 Using Regression Models

South Asia (SEA) is a priority area of environmental pollution and haze conditions. RF validation worked somewhat better than SVR for spatial models, with statistical indicators of $R^2 = 0.76$, RMSE = 11.47 g for urban/industrial sites and $R^2 = 0.64$, RMSE = 10.76 for suburban/rural sites. The goal of this work is to predict $PM_{2.5}$ concentrations in Malaysia utilizing (ML) models derived from satellite AOD (aerosol optical depth) data, floor harmful emissions and weather parameters [5]. SVR calibration performed marginally better than RF calibration for the whole model, with $R^2 = 0.69$ and RMSE = 10.62 versus observed $PM_{2.5}$ concentrations. The author proposed that in future study, gaseous pollutants from satellite remote sensing data be included in ML techniques to estimate $PM_{2.5}$ concentrations.

In recent years, weather and transportation variables, the use of fossil fuels, and industrial characteristics have all played important roles in air pollution. The purpose of this research is to utilise the regression analysis approach to examine the relationship between these variables and predict carbon monoxide CO. Based on other data. The findings reveal that Lasso Regression shows the best fit model. Support environmentalists and the government in developing air quality standards and regulations based on

hazardous and pathogenic air exposure and health-related threats to human welfare. This research may be enhanced by accounting for data changes in real-time prediction over time [6].

For many years, the excessive concentration of particulate matter of size PM10 and $PM_{2.5}$ has had serious health implications. The information used to train the model was obtained from the Taiwan Air Quality Monitoring Network (TAQMN). According to the final results, Gradient Booting regression is the best fitted model. The authors used many regression models, including Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, Gradient Boosting Regressor, and MLP Regressor [7].

In general, rapid growth, urbanisation, and improved living have markedly expanded urban air pollution. A dataset from the Central Pollution Control Board (CPCB) location and an "Air and Noise Pollution Monitoring System" are obtained for the study. Since three stations were evaluated, the R-squared value for Gradient Boosting Regression is 0.69647 in R.K Puram, which is better than other models. However, the scientists advised that other meteorological factors such as precipitation, minimum and maximum temperature, solar radiation, vapour pressure, and so on include it to improve the system's accuracy in future studies.

Malaysia experiences transboundary haze events each year the air includes suspended solids, especially PM10, it has an influence affect man and the environment. The conclusion of the research focuses primarily on study findings that can help responsible parties provide early warning information, as well as mitigation and preventative activities to enhance air quality during haze episodes and human health. The dataset was received from Malaysia's Air Quality Division, Department of Environment (DOE), and Ministry of Natural Resources and Environment in order to conduct more in-depth investigations. As the result shows the multiple linear regression outperformed better compared to other regression models [8].

The increasing frequency of haze weather, the prediction of $PM_{2.5}$ concentrations, the principal pollutant in haze weather, has steadily become a hot subject. The paper's overall goal is to forecast $PM_{2.5}$ concentrations using a multivariate linear regression model. In order to conduct the research, the authors used data from the China Air Quality Online Monitoring and Analysis Platform. The results reveal that R-Squared = 0.8782 and F test = 98.4152, showing that the model is well fitted and can be predicted by the model [9].

Air pollution has been found as a significant predictor of human health. The goal of this project is to forecast pollution with 4 complex regression algorithms, followed by a comparative analysis to determine which model is best for reliably forecasting air quality. Random Forest regression was successful in identifying the peak values. In fact, it takes lesser time to analyse than some other models. For the various data sets, the MAE ranged from 6% to 18%, whereas the RMSE ranged from 0.05 to 0.18. Random Forest regression worked well after hyperparameter adjustment as well [10].

The goal of the present study is to identify the work of other specialists on air pollution and the limitations of their work. The key research gaps found are that just a few studies have been evaluated for the prediction of $PM_{2.5}$ using machine learning. Despite the fact that countless studies on $PM_{2.5}$ have been conducted in various aspects and areas, $PM_{2.5}$ has not yet been introduced into the data science sector.

However, the development of machine learning algorithms is widespread, and more perfect predictions may be contrasted inside the algorithms to get a more robust reaction to developing urbanisation. Therefore, Support Vector Regression, Decision Tree Regression and Multiple Linear Regression are chosen to implement in this research because several studies have proven these machine learning models outperformed than the others.

3 Methodology

This segment will go through how regression models like Support Vector Regression, Multiple Linear Regression, and Decision Tree Regression are implemented to determine the best prediction models for particulate matter $PM_{2.5}$ in smart cities' air quality. The procedure will begin with obtaining data, then go on to data cleaning, feature selection, regression modelling, and finally performance measurements. Figure 1 shows the summarized flowchart of the research.

3.1 Dataset Descriptions

The data for this research is obtained from the Kaggle. The dataset for this research is available in CSV format. The dataset contains data from year 2015 to 2020. Dataset consists of 6236 rows and 16 columns. The taken dataset contains the reading of Benzene, Carbon Dioxide (CO), Ammonia (NH3), Nitric Oxide (NO), Nitrogen Dioxide (NO2), Nitrogen Oxides (NOx), Ozone (O3), Particulate Matter 10 (PM10), Particulate Matter 2.5 ($PM_{2.5}$), Sulphur Dioxide (SO2), Toluene and Xylene of cities in India [11].

It is a technique used in data mining that involves transforming raw data into an understandable data. The data is cleaned through processes such as drop rows with null values, resolving the inconsistencies in the data by changing the data type into integer and adding new variables by splitting the date into a new few columns which contains years, months, days and quarter. Furthermore, categorical variables were subjected to one hot encoding to enhance prediction.

3.2 Feature Selection

Feature selection is a method that automates the selection of those qualities in your data that contribute the most to the prediction variable or output of interest [12].

Many models, particularly linear approaches such as linear and logistic regression, might suffer from the existence of irrelevant properties in your data. For the current study, we employed heatmap correlation to choose the best independent variable that correlates with the dependent variable. Heatmap shows the exact correlation range between 0 and 1 which makes easier to understand and readable. Additionally, heatmaps are graphical depictions where each variable is symbolized by a different color [13]. The most correlated variables to $PM_{2.5}$ are variables such as Benzene, CO, NH_3, NO, NO_2, NO_x, O_3, PM_{10}, SO_2, Toluene and Xylene, Months and Quarter are chosen as those are having best correlation values. Figure 2 shows the heatmap outcome for the dataset.

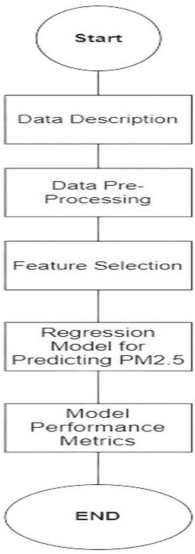

Fig. 1. Summarization flowchart of solution

3.2.1 Support Vector Regression

Support Vector Regression is a supervised learning approach for predicting discrete values. With a few minor modifications, the Support Vector Regression (SVR) employs the same assumptions as the SVM for classification. In the case of regression, a margin of tolerance (epsilon) is supplied to the SVM as an approximate estimate that the problem has already requested [14].

3.2.2 Multiple Linear Regression

Multiple linear regression often known as multiple regression, is a quantitative approach that predicts the result of a response variable using numerous explanatory factors. Linear and non-linear regression use graphs to track a specific answer using two or more variables. Non-linear regression, on the other hand, is typically difficult to implement since it is based on assumptions acquired by trial and error [15].

3.2.3 Decision Tree Regression

The Decision Tree algorithm is a member of the supervised learning algorithm family. The decision tree approach, unlike some other supervised learning methods, may also be utilised to solve regression and classification issues. The purpose of employing a choice tree is to build a training model for predicting the type or quantity of the target attribute by learning basic decision rules from past data (training data). In Decision Trees, we begin at the root of the tree to forecast a target class for a record. The properties of the network of interconnected are calculated and the results of the record's attribute [16].

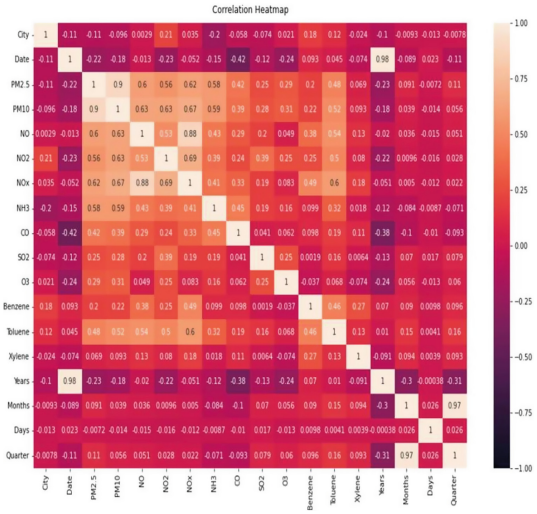

Fig. 2. Heatmap

3.3 Model Performance Metrics

Model evaluation is an essential step in the model creation process. It aids in determining the optimal model to represent our data and how well the chosen model will perform in the future.

3.3.1 Mean Absolute Error

MAE is a ratio of variations between matched observations reflecting the same phenomena. Equation 1 shows the formula for MAE [17].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |yi - xi| \tag{1}$$

MAE = Mean absolute error
yi = Prediction
xi = True value
n = Total number of data points

3.3.2 R-Squared

R-squared is a linear regression model goodness-of-fit measure. This fact illustrates how much of the variance the external variables explain collectively. R-squared is a convenient 0–100% scale that reflects the strength of association among the model and the predictor variables. Equation 2 shows the formula of R- Squared [17].

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2}$$

R^2 = Coefficient of determination
RSS = Sum of squares of residuals
TSS = Total sum of squares

3.3.3 Root Mean Squared Error

The RMSE is a typical means of quantifying the quality of the model's fit in statistical modelling, particularly regression analysis. Equation 3 shows the formula of RMSE [17].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (xi - \dot{x}i)^2}{N}} \tag{3}$$

RMSE = root-mean-square error
i = variable i
N = number of non-missing data points
xi = actual observations time series
$\dot{x}i$ = estimated time series

3.3.4 Adjusted R-Squared

Adjusted R-squared is a version of R-squared that has been modified to adjust for the number of predictors in the model. When the new term improves the model more than would be anticipated by chance, the adjusted R-squared increases. Traditionally, the adjusted R-squared is positive instead of negative. It is always less than R-squared [17]. Equation 4 shows the formula of implementation.

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

R^2 = Sample R-squared
N = Total sample size
P = Number of independent variables

4 Results and Discussion

The precise strategies to determine the best fit model in forecasting $PM_{2.5}$ are thoroughly discussed in this segment, along with performance measures. The segment's flow begins with the performance of Support Vector Regression, then moves on to Multiple Linear Regression, Decision Tree Regression, Adjusted R Squared, and overall performance. To acquire the better performance metrics value, all the regression models were generally undergone hyperparameter tuning using Grid Search method.

Table 1. SVR performance metrics

Without parameter tuning			With parameter tuning		
MAE	RMSE	R^2	MAE	RMSE	R^2
15.36	33.42	0.65	13.10	23.94	0.82

Table 2. MLR performance metrics

Without parameter tuning			With parameter tuning		
MAE	RMSE	R^2	MAE	RMSE	R^2
13.29	23.42	0.82	13.28	23.42	0.83

4.1 Performance of Support Vector Regression

The SVR regressor was used in this experiment to learn and predict $PM_{2.5}$ values. The error levels in MAE and RMSE are (15.36 and 33.42, respectively), according to SVR. This issue is most likely fixed after adjusting the hyperparameters. The MAE and RMSE values in testing sets have decreased to (13.10 and 23.94) in the testing set, respectively. Furthermore, the value for R-Squared shows that there is slight increase from 0.65 to 0.82 after hyperparameter tuning. Table 1 depict the result of performance metrics.

4.2 Performance of Multiple Linear Regression

In addition to SVR, MLR regression was employed to anticipate $PM_{2.5}$ levels. The findings reveal that the MAE and RMSE in the testing set are (13.29 and 23.42). After some parameter adjusting, there is a modest drop in testing, with MAE and RMSE in the testing set showing (13.28 and 23.42). Furthermore, R-Squared results shows there have a higher value after hyperparameter tuning which is from 0.82 to 0.83. The statistics for the performance measures are shown in Table 2.

4.3 Performance of Decision Tree Regression

The final model was implemented as a Decision Tree Regressor. Using the default values, a Decision Tree is utilised to forecast the value of $PM_{2.5}$. The MAE and RMSE results the testing set are 11.55 and 22.24. This error was reduced after some hyperparameter adjustment, yielding an MAE of 10.80 and an RMSE of 20.70 in the testing set. In addition, corrected R-Squared was calculated using variable interactions. R-Squared results shows 0.84 before the hyperparameter tuning while 0.86 after the tuning in hyperparameter. The result for the performance measures is shown in Table 3.

4.4 Adjusted R-Squared

The R^2 and adjusted R^2 of three models are compared in Table 4. It illustrates that the Decision Tree's R^2 value is the best to choose because it has the highest value. Despite

Table 3. DTR performance metrics

Without parameter tuning			With parameter tuning		
MAE	RMSE	R^2	MAE	RMSE	R^2
11.55	22.24	0.84	10.80	20.70	0.86

Table 4. Comparison of R^2 and adjusted R^2

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

having the maximum value, R^2 may lead to overfitting by the chosen predictors which is 13. By varying the number of independent variables, the modified R-Squared is utilised to avoid overfitting. It works by selecting only limited number of independent variables. The greatest number of independent variables that may be investigated using adjusted R^2 in this case is 13. As a result, as the overall research, 4 variables will be picked since the values for 8 and 13 drop significantly, indicating that the added variables may not correlate to the target variable.

4.5 Overall Performance

Figure 3 depicts the overall performance metrics of three models in bar chart style, while Table 5 displays the overall RMSE, R2 and MAE values of the three models.

5 Conclusion and Future Work

Due to the extent of air pollution, machine learning models were evaluated in this study for estimating Air Pollution in $PM_{2.5}$ in urban areas. Therefore, Decision Tree Regression surpassed SVR and MLR by showing the better prediction with the lowest error in RMSE which is 20.7084 and higher value in R-Squared 0.8666 after hyperparameter tuning. Furthermore, since 4 variables were chosen for adjusted R^2, the Decision Tree value of 0.8620 outperforms MLR and SVR. Future work may be done by experimenting with other variables that impact air pollution. Furthermore, it would be preferable if more accurate datasets and values were fully completed in order to provide more accurate predictions in the future.

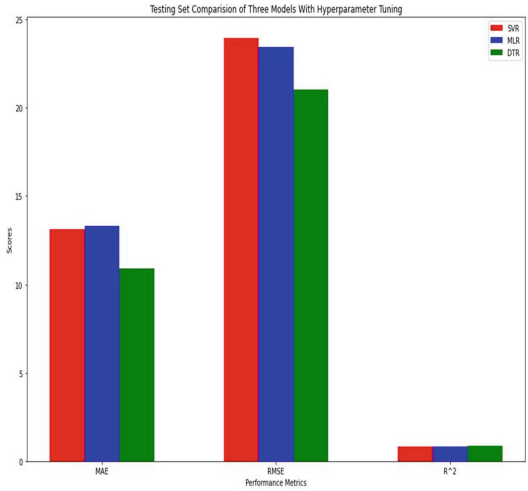

Fig. 3. Overall performance metrics of three models

Table 5. Overall performance metrics after hyperparameter tuning

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

References

1. S. Shea and E. Burns, "Smart city," TechTarget, 2020. [Online]. Available: https://www.techtarget.com/iotagenda/definition/smart-city.
2. "Air pollution," World Health Organization, 2021. [Online]. Available: https://www.who.int/health-topics/air-pollution#tab=tab_1.
3. Ameer S, Shah MA, Khan A, Song H, Maple C, Islam SU, Asghar MN. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access, 2019, pp.128325–128338.
4. L. Teeboom, "The Advantages of Regression Analysis & Forecasting," 8 3 2019. [Online]. Available: https://smallbusiness.chron.com/advantages-regression-analysis-forecasting-61800.html.
5. Zaman NA, Kanniah KD, Kaskaoutis DG, Latif MT. Evaluation of Machine Learning Models for Estimating $PM_{2.5}$ Concentrations across Malaysia. Applied Sciences, 2021, pp.7326.
6. Abdullah S, Napi NN, Ahmed AN, Mansor WN, Mansor AA, Ismail M, Abdullah AM, Ramly ZT. Development of multiple linear regression for particulate matter (PM_{10}) forecasting during episodic transboundary haze event in Malaysia. Atmosphere, 2020, p. 289.
7. Harishkumar KS, Yogesh KM, Gad I. Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models. Procedia Computer Science, 2020, pp. 2057–2066.

8. Abdullah S, Napi NN, Ahmed AN, Mansor WN, Mansor AA, Ismail M, Abdullah AM, Ramly ZT. Development of multiple linear regression for particulate matter (PM_{10}) forecasting during episodic transboundary haze event in Malaysia. Atmosphere, 2020, pp.289.
9. Chen J, Wang J. Prediction of $PM_{2.5}$ concentration based on multiple linear regression. In2019 International Conference on Smart Grid and Electrical Automation (ICSGEA), 2019, pp. 457–460, IEEE.
10. Srivastava C, Singh S, Singh AP. 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.
11. H. Vora, "city_day.csv," 2020. [Online]. Available: https://www.kaggle.com/datasets/hirenvora/city-daycsv?resource=download.
12. R. Shaikh, "Feature Selection Techniques in Machine Learning with Python," 2018. [Online]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e.
13. A. Kumar, "Correlation Concepts, Matrix & Heatmap using Seaborn," 2022. [Online]. Available: https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/.
14. A. Sethi , "Support Vector Regression Tutorial for Machine Learning," 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/.
15. R. Bevans, "Multiple Linear Regression | A Quick Guide (Examples)," Scribbr, 2020. [Online]. Available: https://www.scribbr.com/statistics/multiple-linear-regression/.
16. N. S. Chauhan, "Decision Tree Algorithm, Explained," KD nugget, 2022. [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html.
17. S. Hiregoudar, "Ways to Evaluate Regression Models," 2020. [Online]. Available: https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70.