IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Estimating of Particulate Matter using Machine Learning Approaches

Jash J Patel[1], Prof. Neha R Patel[2], Prof. Hetal Gaudani[3]

[1:] M. Tech. Environmental Engineering Student, Department of Civil Engineering, BVM Engineering College, Vallabh Vidyanagar,
Gujarat, India.

[2:] Assistant Professor, Department of Civil Engineering, BVM Engineering College, Vallabh Vidyanagar, Gujarat, India.

[3:] Assistant Professor, Department of Computer Engineering, GCET Engineering College, Vallabh Vidhyanagar, Gujarat, India

Abstract: The main causes of air pollution in India's metropolitan areas have been linked to urbanization and industrial expansion. Air pollution has been identified as one of the most serious issues in metropolitan areas. Particulate matter is still one of the most significant sources of air pollution in cities, and both acute and chronic exposures have major health consequences. Meteorological factors play a key role in determining the presence of particulate matter. As a result, the primary goal of this study is to examine and assess machine learning models and methods that deal with a specific field, namely Particulate Matter, as well as to determine the strengths of various machine learning techniques and give background information. This study also intends to emphasize the role of input predictors in enhancing predictive accuracy and the core methodologies of machine learning and their importance in enhancing prediction performance. This study reveals that how successful it is to combine machine learning approaches with Particulate Matter prediction.

Keywords: **Particulate Matter, Meteorological factors, Random Forest, Support Vector Machines and Prediction.**

1. Introduction

Every living organism on earth is dependent on air as one of its major components. There has been an increase in pollution over the last 50 years due to urbanization, industrialization, automobiles, power plants, and chemical activity as well as other natural activities such as volcanic eruptions, agricultural burning, and wildfires. All these activities cause pollution growth, particularly particulate matter (PM) is one of the significant reasons for air pollution (Jung, 2017). Pollution is caused by several factors, including stubble burning along with hazardous particulates such as PM2.5 and PM10 (Zanobetti et al. 2009). In general, these particulate matters are composed primarily of solids and liquids suspended in the air, and they have diverse chemical compositions, including some organic compounds, sulfur dioxide, etc (Davidson, 2005). These particles are primarily composed of PM2.5 particles, which, as their name implies, refer to fine atmospheric particulate matter with a diameter smaller than 2.5 m", which is about 3% the diameter of a human hair. Particles like these are extremely hazardous for health because they can easily penetrate deep into the lungs, irritate the alveolar wall, and corrode it. All of this results in the lungs being severely impaired. There are numerous negative effects of PM2.5, not only asthma, respiratory inflammation, cardiovascular diseases, but even cancer may be caused by it (Valavanidis et al. 2006). These fine patches, if entered into the lungs, might round the inflexibility of COVID-19 infection as the new coronavirus also attacks the respiratory system (Kumar et al. 2020). If the attention of these pollutant patches is veritably high in the atmosphere, it oppressively affects our health and may beget life-changing problems in a short period (Graff, 2007). Studies have established that particulate matters affect mortal health indeed at the inheritable position. This paper aims to find the best machine learning prediction model like Random Forest, Gradient Boost, Decision Tree, and Regression models.

1.1 Particulate Matter

Particulate air pollution is a suspension in the air of a mixture of solid, liquid, or solid and liquid particles. The size, nature, and origin of these dispersed particles vary. Aerodynamic qualities are useful for classifying particles because they influence particle transit and removal from the air, as well as particle deposition inside the respiratory system, and they are linked to particle chemical composition and sources. Particulate matter (PM) is a well-known indoor and outdoor air contaminant that ranges in size from a few nanometres to tens of micrometers. PM in ambient air originates from natural sources, anthropogenic sources, and atmospheric transformation. The main sources of indoor particulate matter include penetration from outside air, cooking, and resuspension from home dust (Pope et al. 2020). Under some conditions, indoor air chemistry could also be a significant contributor to indoor PM. PM may comprise hundreds of inorganic and organic species, despite being controlled for bulk as a single chemical. The size and

chemical content of PM vary greatly depending on the source. Mechanical processes, such as resuspended road dust, abrasive mechanical operations in industry and agriculture, and some bioaerosols, are the primary sources of coarse PM (particles having a diameter of 2.5–10 mm). PM2.5, particles with size ranges from 0.1 to 2.5 mm, and ultrafine particles PM10, 10mm particles with size. (Z. Fan et al. 2008)

1.2 Machine Learning

The study of computer algorithms that can learn and evolve on their own given experience and data is known as machine learning. It's thought to be a part of artificial intelligence. Machine learning algorithms create a model from training data and use it to make predictions or make judgments without the need for explicit programming. The steps in analyzing a machine learning model are as follows: Data Understanding- It is necessary to first assess the raw data before constructing the various ways of working with data. Data preparation can include things like merging data, imputing missing values, deleting variables with too many missing values, sorting data, and so on. Model training is the process of putting models through their paces and evaluating the results. The evaluation of the results is an important step in interpreting the findings since it allows the models to be tweaked and the initial research method to be tweaked. In addition, describing and illustrating the various models that can be used: The process of supervised learning entails matching a set of independent factors to one or more dependent variables. Examples of these types of tasks include regression and classification. Unsupervised learning, on the other hand, does not require any prior "correct" data, and the purpose of this form of research is to uncover the data's underlying patterns. Optimization strategies are methods for discovering the best set of parameters to minimize a pre-defined cost function. (Wilcox et al. 2013).

1.2.1 Decision Tree

A decision tree is a graph that represents choices and their outcomes as a tree. The graph's nodes represent events or choices, while the graph's edges reflect decision rules or conditions. Nodes and branches constitute each tree. Each branch indicates a value that the node can take, and each node represents attributes in a group that needs to be categorized.

J Wang et al. (2015) used PM2.5 concentration data as well as meteorological data from January 1 to December 31, 2013. They selected the Nagasaki region of Japan for this study. And their findings of the correlation analysis between PM2.5 concentration and meteorological data revealed that temperature had a negative association with PM2.5, whereas precipitation had a positive correlation. The correlations between humidity and wind speed and PM2.5 had a threshold, and they discovered that depending on whether the meteorological variable values were lower or higher than the threshold, the correlation was positive or negative. And found that the west wind may deliver the greatest pollutants to Nagasaki, Japan, based on the association between wind direction and pollution.

For this study, **Y Gao, (2021)** collected data from 20 monitoring points and an experimental site on a neighborhood scale of 2 km*2 km in the Minhang District of Shanghai, China. This study analyzed PM2.5 and O3 concentrations and meteorological parameters such as solar radiation, relative humidity, air temperature, and green space. In this study, they used the Decision Tree approach and discovered that a decision tree model enhanced the accuracy, effectiveness, and time resolution of predicting the spatial variation of air pollutants by 14 percent–21%. And they conclude that this work illustrates the superiority of decision tree models in simulating spatial fluctuations of O3 and PM2.5 concentrations at a neighborhood scale, creating an opportunity for further research in this area.

For this work, **T Zhang et al. (2020)** used the Gradient Boosting Decision Tree model to estimate PM2.5 concentrations, and they obtained daily average PM2.5 readings from the China National Environmental Monitoring Center and local environmental monitoring centers in 2017. Furthermore, they combined various independent factors to construct a Gradient Boosting Decision Tree model to predict daily ground PM2.5 concentrations at a 3-km spatial resolution across China using a Linear Regression model to fill in missing satellite aerosol optical depth data and They discovered that the Gradient Boosting Decision Tree model worked well in estimating temporal variability and geographic differences in daily PM2.5 concentrations, with 0.98 fitted model coefficients of determination, 3.82g/m3 root mean square errors, and 1.44 g/m3 mean absolute error. They conclude that this method may be used to increase the accuracy of PM2.5 estimation with higher spatial resolution, particularly in the summer, and It also can be used to improve the precision of ground-based satellite-based PM2.5 estimation.

R. Waman et al. (2017) provide a system for categorizing the health hazards of air pollutants based on AQI criteria and emphasizing air quality based on data from various air pollutants like NO_2, SO_2, CO, and O_3. To forecast the health condition, their research uses the Naive Bayes method and the Decision Tree algorithm. The Air Quality Index is divided into four categories: good, moderate, unhealthy, and very unhealthy. Air Quality Index standards are used to classify the health risks of air contaminants, and the classifiers in this study are: The level of risk for Air Quality Index values in the range of 0 to 50 is "GOOD," 51 to 100 is "MODERATE," 101 to 150 is "unhealthy for sensitive populations," 151 to 200 is "UNHEALTHY," 201 to 300 is "VERY UNHEALTHY," and over 300 is "VERY DANGEROUS.". And the result shows that the decision tree algorithm Provides an accuracy of 91.9978% which is more the algorithm of Naïve Bayes.

In their research, **Y. Rybarczyk et al. (2016)** tackled the problem of forecasting fine particulate matter (PM2.5) in light of a combination of weather circumstances. Several years of meteorological data and a machine learning method were utilized to develop a model in Quito, Ecuador. In this study, a Decision Tree approach is used to categorize concentrations into two classes (> 15g / m3 vs. 15g / m3) using a minimal number of parameters such as precipitation amount, wind speed, and wind direction. Models generated

from a few rules can accurately anticipate the concentration result. The classification results obtained with the decision tree are compared to those obtained with other classifiers to see whether there are any significant differences in classification between all of the classifier models. In comparison to other classification models, they found that Decision Tree predicts PM2.5 concentrations based on a threshold value of 15g/m3 and a comparatively high proportion of successfully categorized instances over 65%.

1.2.2 Random Forest

Random Forest Regression is a supervised learning approach for regression that uses the ensemble learning method. The ensemble learning method combines predictions from several machine learning algorithms to get a more accurate forecast than a single model. Random Forest Regression is a powerful and precise model. It usually works well in a wide range of situations, including those with non-linear relationships. However, there are some drawbacks: there is no interpretability, overfitting is a possibility, and we must choose the number of trees to include in the model. These two techniques are used. 1. Boosting technique: The term "boosting" refers to a group of algorithms that help weak learners become stronger. Boosting is a bias and variance reduction strategy used in ensemble learning. A classifier is defined as a weak learner, while a strong learner is defined as a classifier that is arbitrarily well-correlated with the real classification. 2. Bagging technique: When the accuracy and stability of a machine learning algorithm need to be improved, bagging or bootstrap aggregating is used. Bagging also reduces variation and aids in the management of overfitting.

I Kloog et al. (2019) developed an ensemble model that included different machine learning techniques and predicted values to estimate daily PM2.5 with a resolution of 1 km × 1 km over the contiguous United States. They gathered meteorological data from NOAA's North American Regional Reanalysis data sets, which were obtained between January 1st, 2000, and December 31st, 2015. Satellite observations, land-use terms, climatic data, and other predictor variables were used in this. The mean R^2 between daily predicted and measured PM2.5 after cross-validation was 0.86, with an RMSE of 2.79 g/m3. At the yearly level, R^2 was 0.89, suggesting strong model performance. They discovered that a single machine learning algorithm may underperform at a given year, season, location, pollution concentration, and so on, and an ensemble model including estimation from numerous machine learning algorithms might achieve improved model performance.

The goal of this work by **B. Bashir et al. (2019)** was to determine the significance of features for PM2.5 prediction. The location they chose was Tehran, Iran's capital, and the data-gathering period was from 2015 to 2018. They use the random forest, extreme gradient boosting, and deep learning machine learning methodologies to determine the feature impact for PM2.5 prediction in Tehran's metropolitan region. As a result, XGBoost outperformed Random Forest and Deep Learning algorithms with R2 = 0.81, R = 0.9, MAE = 09.92 g/m3, and Root Mean Square Error = 13.58 g/m3 at a very low cost of 19 s. Although a DNN model was employed for modeling and prediction, XGBoost performed better because of its basic structure. As a result, they demonstrated that this model is superior to the other models studied. Decision trees, which belong under the domain of supervised ensemble learning techniques, give rise to the random forest.

In this research, **S Singh et al. (2020)** suggested a successful Machine Learning-based model for the prediction of PM2.5 as an air quality metric in Delhi's atmosphere and used the Extra Tree Regression and Adaptive Boosting are combined in the proposed machine learning-based approach. They compared their results to those of other existing models such as Random Trees, Decision Trees, and so on. They use the Mean Absolute Error, Root Mean Square Error, and R^2 score as performance indicators to compare their model to other current models. The findings demonstrate that the ET+AdaBoost performs better than other models, with R2=92.64, MSE=14.79, and RMSE=25.11.

For this study, **C Johansson et al. (2020)** proposed a random forest model for prediction. PM10, PM2.5, NO_2, and O_3 were the parameters they used for each square kilometer of Sweden from 2005 to 2016. For this study, the Air Quality data were taken from the Swedish Meteorological and Hydrological Institute. The study's main aim is to create a regular grid with a 1-km2 resolution over Sweden. As predictors of air pollution variability over time and place, they incorporated satellite data, atmospheric composition factors, land-use terms, meteorological parameters, and population density. With cross-validated R2 in the range of 0.64–0.77 for out-of-bag samples and 0.37–0.60 for held-out monitors, their models showed no bias and were able to predict the majority of the variability. They were successful in creating the regular grid.

C. Feng et al. (2017) proposed a fine-grained PM2.5 estimating approach based on the random forest algorithm without any PM2.5 measuring devices. The five data sources they use to evaluate their work include meteorological and traffic data, records from monitoring sites, POIs, and lastly the images they take. The results of the study were compared to other methods and showed that when the random forest method was used to estimate PM2.5, it had a high level of accuracy (the precision was 87.5 percent and the recall was 87.2 percent), which was superior to the other methods Logistic, Nave Bayes, Random Tree, and BP ANN.

1.2.3 Support Vector Machine

Another popular state-of-the-art machine learning approach is the Support Vector Machine. In machine learning, support vector machines, or SVMs, are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. By implicitly mapping their inputs into high-dimensional feature spaces, SVMs may do both non-linear and linear classification. This is known as the kernel trick. It's utilized to make distinctions between classes. The margins are drawn so that there is a little gap between the margin and the classes as feasible, which reduces the classification error.

J Deters et al. (2017) proposed a machine learning strategy for predicting PM2.5 concentrations from wind speed, wind direction, and precipitation levels based on six years of meteorological and pollution data studies. They used two north-western air quality monitoring stations for their study: Cotocollao and Belisario District. In this study, they represent a machine learning strategy for predicting PM2.5 concentrations in a high-elevation mid-sized city using meteorological data. They employed Boosted Trees, linear support vector machine, neural network, and convolutional generalization model approaches to predict PM2.5 levels. And the AUC-BT (0.72) value is higher than the AUC-L-SVM (0.66) value, with an MSE of 22.1 percent for NN and 15.6 percent for CGM. As a result, the Regression analysis reveals that when climatic circumstances become more extreme, a better prediction of PM2.5 may be made and it is supported by the strong correlation between estimated and real data for a time series analysis during the wet season. Their study demonstrates that statistical models based on machine learning are relevant for predicting PM2.5 concentrations from meteorological data.

Masood et al. (2020) concentrated on the Delhi area in their article. They developed several machine learning approaches that were used to forecast daily PM2.5 concentrations in Delhi. On the inputs of various meteorological and pollution characteristics corresponding to two years from 2016 to 2018, two different models, Support Vector Machine and Artificial Neural Networks, were created. They discovered that for the available test dataset, the Artificial Neural Networks-based PM2.5 prediction model integrating air pollution and meteorological data performed better. As a consequence, they discovered that Artificial Neural Networks can outperform conventional machine learning approaches in this study when it comes to predicting PM2.5 concentrations. The training and testing correlation values for the Artificial Neural Networks model were found to be 0.856 and 0.730, respectively, indicating the model's appropriateness for PM2.5 prediction. As a result, the Artificial Neural Networks model with improved generalization capabilities may be considered an optimum alternative approach for model building for multi-dimensional complicated situations such as air pollution.

Z. Qin et al. (2017) gathered particulate samples from Shenzhen, China's south coast, and developed a new hybrid-Garch Generalized Autoregressive Conditional Heteroskedasticity approach to merge the Autoregressive Integrated Moving Average and forecasting models Support Vector Machine. Data from 10-day hourly PM2.5 concentrations, both linear and non-linear, are used to evaluate the hybrid arch technique for time series prediction. The PM2.5 concentrations in Shenzhen have a regular variation throughout the 24 hours of the day, with the greatest value during working hours due to plant and vehicle emissions, according to empirical results from six-station data sets. Because of the geographical and climatic circumstances, the spatial variation in PM2.5 concentrations is not visible. They suggested hybrid model produces a more precise and dependable forecast and they also suggested hybrid model examines time-series data that may exhibit conditional error variance. and estimates the variance for the volatility of the PM2.5 concentrations. And as a result, the Support Vector Machine performs better than another model.

For this study**, S. Revathy et al. (2021)** used six basic machine learning algorithms: logistic regression, decision tree, support vector machine, random forest tree, Nave Bayes theorem, and K-nearest neighbor, and used PM10, PM2.5, SO_2, CO, and NO_2 parameters for the prediction of the Air Quality Index for Delhi. As a consequence, they discovered that the decision tree, which has a precision of 99.88 percent, is the most successful approach, while the Support Vector Machine Classifier, which has a precision of 70.65 percent, is the least precise algorithm. The Random Forest, on the other hand, has an accuracy of 99.16 percent, which is practically identical to that of the Decision Tree. They discovered that the random forest is a type of decision tree and has nearly the same accuracy as a decision tree. They also discovered that logistic regression, Nave Bayes, and K-nearest neighbor all have an accuracy of around 97 percent.

1.2.4 Naive Bayes

Naive Bayes is a classification method based on the Bayes Theorem and the assumption of predictor independence. A Naive Bayes classifier, in simple terms, assumes that the existence of one feature in a class is unrelated to the presence of any other feature. The text classification industry is the primary focus of Naive Bayes. It's primarily used for grouping and classification, and it's based on the conditional probability of occurrence.

Rubal et al. (2018) conducted their research using Random Forest, Naive Bayes, and Multilayer Classifier models. For the prediction, they selected the cities of Delhi and Patna. C_6H_6, NO_2, CO, SO_2, O_3, $PM_{2.5}$, and PM_{10} were the pollutants used in this investigation. From the Central Pollution Control Board, they anticipated pollutants such as C_6H_6, NO_2, and CO from Delhi, and SO_2, O_3, $PM_{2.5}$, and PM_{10} from Patna. As a result, they suggest that combining a differential evolution strategy with the random forest method outperforms using the Bayesian network's independent classifier and the multi-label classifier methodology separately.

Dejan Petelin et al. (2013) proposed the Naive Bayes approach for this study, and they used the Bourgas, Bulgaria region for prediction. For this investigation, ozone, sulfur dioxide, nitrogen dioxide, phenol, and benzene concentrations, as well as meteorological factors such as wind speed, wind direction, and temperature, were used. They find that the Naive Bayes technique for predicting ozone concentrations appears promising. The influence of exponential forgetting on ozone concentration forecasts is examined, and it is discovered that the forgetting factor k = 0.99985 yields the most accurate predictions. they also conclude that this strategy is required when training data is not available for the entire period of interest, resulting in the inability to learn all period characteristics, or when the environment to be modeled is constantly changing.

R. Waman et al. (2017) provide a system for categorizing the health hazards of air pollutants based on Air Quality Index criteria and emphasizing air quality based on data from various air pollutants (NO_2, SO_2, CO, and O_3). To forecast the health condition, their research uses the Naive Bayes method and the Decision Tree algorithm. The Air Quality Index is divided into four categories: good, moderate, unhealthy, and very unhealthy. Air Quality Index standards are used to classify the health risks of air contaminants, and the classifiers in this study are: The level of risk for Air Quality Index values in the range of 0 to 50 is "GOOD," 51 to 100 is "MODERATE," 101 to 150 is "unhealthy for sensitive populations," 151 to 200 is "UNHEALTHY," 201 to 300 is "VERY UNHEALTHY," and over 300 it is "VERY HARMFUL.". And the result shows that the decision tree algorithm Provides an accuracy of 91.9978% which is more than the algorithm of Naive Bayes. And Naive Bayes method was not accurate as compared to the decision tree.

1.2.5 Neural Networks

A neural network is a collection of algorithms that attempt to recognize underlying relationships in a set of data by simulating how the human brain works. Neural networks, in this context, are biological or artificial systems of neurons. Because neural networks adjust to changing input, they can deliver the best possible outcome without rethinking the output criteria.

H Kumar et al. (2020) has developed a Machine Learning-based model for predicting PM2.5 as an air quality metric in Taiwan's atmosphere. They gathered data from the Air Quality Monitoring system between 2012 and 2017 and they conducted comparison research using performance indicators such as Random Forest, Decision Trees, Gradient Boosting Regression, and Multiple Linear Regression and as Machine learning algorithms based on statistical estimates of metrics such as Mean Absolute Error, Mean Square Error, Root Mean Squared Error, and Coefficient of determination were used to forecast particulate matter PM2.5. And their findings demonstrate that the suggested model's values perform better than the previous models, with R^2 =0.89, MSE=0.0619, RMSE=0.1302, and MAE=0.0380, indicating that the actual and predicted values are quite similar to each other. It concludes that the gradient boosting regressor model is better for forecasting air pollution on the Taiwan Air Quality Network data.

In this study, **X Feng et al. (2015)** used 13 distinct air pollution monitoring sites of China's northeastern cities to create the model. In addition, they prepared a unique hybrid model that was presented to forecast daily average PM2.5 concentrations two days ahead of time by combining a trajectory-based geographic model with a wavelet transformation in a Multiple Linear Regression type of neural network and when this hybrid model, when combined with meteorological predictions and pollutant predictors, is thought to be an efficient method for improving PM2.5 forecasting accuracy. As a consequence, they have proven that the trajectory-based geographic model and wavelet processing have been useful techniques for improving PM2.5 forecasting accuracy, with the hybrid model's root mean squared error decreased by up to 40% on average. High PM2.5 days, in particular, may practically be predicted using this method. As a result, the method they present here may be applied to various places and produces better predicting accuracy.

W You et al. (2016) created a countrywide, geographically weighted regression model to predict ground-level PM2.5 concentrations in China using recently disclosed nationwide, hourly PM2.5 values for this study. They took primary predictor of a 3 km resolution aerosol optical depth output from the Moderate Resolution Imaging Spectroradiometer. And their findings of the geographically weighted regression model's performance suggested that it was quite accurate in estimating ground-level PM2.5 concentrations. With a Root Mean Square Error of 18.6 g/m3, the geographically weighted regression model was able to explain almost 79 percent of the variability in daily PM2.5 concentrations, and these findings are valuable for health risk assessment, air pollution management measures, and environmental research. The findings also revealed that the geographically weighted regression model used in this study is capable of detecting PM2.5 spatial patterns at various scales. The findings of mapping national-scale PM2.5 concentrations can also be utilized to help China's future monitoring building plans.

A. Suleiman et al. (2016) investigate how Artificial Neural Networks and Boosted Regression Tree approaches can be used to model air quality. Based on air pollution, traffic, and meteorological data, the approaches were used to create air quality models for predicting roadside particle mass concentration (PM10, PM2.5) and particle number counts. They've chosen the Marylebone Road area of London for this research. They compared Fraction of predictions within a factor of two of the observations, mean bias, Mean Squared Error, Normalised Mean Bias, Root mean square error, R, and Coefficient of efficiency values to check the prediction accuracy of the Artificial Neural Network and Boosted Regression Tree models, and found that 87– 99 percent of the model predictions are within a factor of two of the observed data, indicating good agreement between the model predictions and particle observations. And they found that the models' Coefficient of efficiency values range from 0.70 to 0.81, indicating that they can forecast particle concentrations substantially more accurately than the mean of observed concentrations. Conclusion and conclude that the artificial neural networks models were marginally better than the Boosted regression tree models in terms of accuracy.

S. Malewar, (2020) in his research paper Improving Neural Network Prediction Accuracy for PM_{10} Individual Air Quality Index Pollution Levels stressed pollutants having a diameter less than <10 µm (PM_{10}) in two major cities of China. The reason for the generation of fugitive dust was due to construction activities and was interlinked with Construction Influence Index. His Neural Network Models were based on perceptron, Elman, and Support Vector Machine. The dataset was decomposed into wavelet representations and then wavelet representations were predicted. His predictions were tested between 1 January 2005, and 31 December 2011, at six monitoring stations situated within the urban area of the city of Wuhan, China. It yielded better results than previous models but he only focussed on pollutants. And as a result, they were successful in Neural Network Prediction Accuracy for PM_{10} Individual Air Quality Index Pollution.

1.2.6 Regression Model

P. Goyal et al. (2011) published a study that used Principal Component Regression and Multiple Linear Regression Techniques to anticipate the daily Air Quality Index value for the city of Delhi, India, utilizing past Air Quality Index and meteorological parameters. They use prior records from the years 2000 to 2005 and various formulae to predict the daily Air Quality Index for the year 2006. In this PM_{10}, $PM_{2.5}$, SO_2, CO, and NO_2 parameters for the prediction of the Air Quality Index Then, using the Multiple Linear Regression Technique, this predicted value was compared to the observed value of Air Quality Index in 2006 for the season's summer, monsoon, post-monsoon, and winter. The collinearity between the independent variables is determined using Principal Component Analysis. Multiple Linear Regression employed principal components to remove collinearity among predictor variables and reduce the number of predictors. In comparison to other seasons, the Principal Component Regression performs better in forecasting the Air Quality Index in the winter.

In their paper, **O. Kisi et al. (2017)** explore the forecasting of SO_2 concentration using three different soft computing approaches least square support vector regression (LSSVR), multivariate adaptive regression splines (MARS), and M5 model tree. All of the models are applied to data collected every month in Delhi, India, Nizamuddin, Janakpur, and Shahzadabad. All of the models are compared and evaluated using the root mean square error, mean absolute error, and correlation coefficient. Based on the results of the comparison, least-square support vector regression outperformed all other models in terms of accuracy, while the MARS model was ranked second in terms of SO_2 prediction. According to their findings, all models improved the Janakpur station's forecasting accuracy.

A Chaloulakou et al. (2003) research has implemented Artificial Neural Network and Multiple Linear Regression algorithms to forecast the PM10 concentration over the two years for the city of Athens, Greece. Before applying an input to Artificial Neural Network, the dataset is divided into three unequal subsets as the training dataset contains two-third of the available records or cases and the remaining cases were equally divided into validation and test set. Comparison between Artificial Neural Network and Multiple Linear Regression was also done in this study that indicates Multiple Linear Regression is better in performance than Artificial Neural Network. According to this study, Multiple Linear Regression will give adequate prediction solutions or results as per the requirement if it is properly trained.

Nidhi Sharma et al. (2018) examined thorough data on air pollutants from 2009 to 2017 and offered a critical analysis of the 2016-2017 air pollution trend in Delhi, India. Sulfur Dioxide, Nitrogen Dioxide, Suspended Particulate Matter, Ozone, Carbon Monoxide, and Benzene are among the pollutants for which they have predicted future trends. They predicted the future values of the pollutants mentioned earlier based on previous records using data analytics Time-series Regression forecasting. The monitoring stations of AnandVihar and Shadipur in Delhi are being investigated based on the findings of this study. The results demonstrate a significant increase in PM_{10} concentrations and increases in NO_2 and $PM_{2.5}$, indicating increased pollution in Delhi. CO levels are expected to drop by 0.169 mg/m3, but NO_2 levels are expected to rise by 16.77 mg/m3 in the future years. Ozone levels are expected to rise by 6.11 mg/m3, benzene levels should decrease by 1.33 mg/m3, and SO2 levels will rise by 1.24 mg/m3.

No.	Approach	Region	Data collection duration	Parameters	Algorithms	Best Model as a result of the comparison
1.	Prediction	North American Region, USA	January 1, 2000, to December 31, 2015	$PM_{2.5}$ and Meteorological Data	Random Forest	Random Forest
2.	Prediction	Tehran	2015 to 2018	$PM_{2.5}$ and Meteorological Data	Random Forest, Extreme Gradient Boosting	Random forest
3.	Prediction	Delhi	2017 to 2019	$PM_{2.5}$	Extra Tree Regression, Adaptive Boosting Random Forest	Random Forest
4.	Prediction	Sweden	2005 to 2016	$PM_{2.5}$, PM_{10}, NO, and O_3	Random Forest Regression	Random Forest Regression
5.	Prediction	China	2015 to 2018	Meteorological data, Traffic data, POIs, and Images	Random Forest, Logistic Regression, Decision Tree and Artificial neural network	Random Forest

No.	Approach	Region	Data collection duration	Parameters	Algorithms	Best Model as a result of the comparison
6.	Prediction	Nagasaki, Japan	January 1 to December 31, 2013	$PM_{2.5,}$ Humidity and Wind speed	Random Forest and Decision Tree	Decision Tree
7.	Prediction	Shanghai, China	2012 to 2017	$PM_{2.5,}$ O_3 and Meteorological Factors	Decision Tree	Decision Tree
8.	Prediction	China	2013 to 2016	$PM_{2.5,}$ NO, and O_3	Linear Regression and Gradient Boosting Decision Tree	Gradient Boosting Decision Tree
9.	Forecasting	Taiwan	2012 to 2017	$PM_{2.5}$, SO_2, NO_2, CO, Wind speed, Temperature	RF, GBR, DTR, MLP	Gradient Boosting Decision Tree
10.	Prediction	Quito, Ecuador	2009 to 2015	$PM_{2.5,}$ Wind Speed and Wind Direction	Random Forest, Logistic Regression, Decision Tree and Artificial neural network	Decision Tree

No.	Approach	Region	Data collection duration	Parameters	Algorithms	Best Model as a result of the comparison
11.	Prediction	Delhi	2016 to 2018	$PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO	Support Vector Machine and Artificial Neural Network	Support Vector Machine
12.	Prediction	Shenzhen, China	2007 to 2017	$PM_{2.5}$, O_3 and Meteorological Factors	Linear Regression and Gradient Boosting Decision Tree, Support Vector Machine	Support Vector Machine
13.	Prediction	Delhi	2013 to 2019	$PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO	Logistic Regression, Random Forest, Naïve Bayes, Support Vector and K-nearest neighbor	Support Vector Machine
14.	Prediction	Tehran	2006 to 2016	wind direction, temperature, etc	Support Vector Machine, Artificial Neural Network	Support vector, Machine
15.	Forecasting	Taiwan	2012 to 2017	Co, SO_2, NO_2, CO_2, Wind speed, and Temperature	Random Forest, Gradient Boosting Regressor, DTR, Multilayer Perceptron, and Support Vector Machine	Gradient Boosting Regressor

No.	Approach	Region	Data collection duration	Parameters	Algorithms	Best Model as a result of the comparison
16.	Prediction	Delhi	2016 to 2018	$PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO	Support Vector Machine and Artificial Neural Network	Support Vector Machine
17.	Forecast	Northeastern cities of China	2009 to 2013	$PM_{2.5}$, O_3 and Meteorological Factors	Multiple Linear Regression and Neural Network	Artificial Neural Network
18.	Prediction	China	2012 to 2015	$PM_{2.5}$, PM_{10}, SO_2, and NO_2	Logistic Regression, Random Forest, Support Vector and Neural Network	Artificial Neural Network
19.	Prediction	Marylebone, London	2011 to 2014	$PM_{2.5}$ and PM_{10},	Artificial Neural Network and Boosted Regression Tree	Artificial Neural Network
20.	Prediction	Wuhan, China	1st January 2005 to 31st December 2011	PM_{10}	Artificial Neural Network and Support Vector Machine	Simulated Neural Network

No.	Approach	Region	Data collection duration	Parameters	Algorithms	Best Model as a result of the comparison
21.	Prediction	Delhi	2000 to 2005	$PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO	Regression and Multi Linear Regression	Multi Linear Regression
22.	Prediction	Delhi	1st January 2018 -- 30th November 2019	Wind speed, atmospheric Temperature, Pressure, etc.	Regression model: Extra-Trees regression and AdaBoost	Regression Model
23.	Prediction	Beijing	2015	AOD, Meteorological factors Gaseous pollutants	Multi Linear Regression	Multi Linear Regression
24.	Predicting	Quito, Ecuador, Cotocollao Belisario	2007 - 2013	Aerosol data, fine particle concentrations, meteorological data	Boosted Trees, L-Support Vector Machine, Neural Network and	Multi Linear Regression
25.	Prediction	Athens, Greece	2000 to 2003	Co, SO2, NO2, CO2, Wind speed, Temperature	Artificial Neural Network, Multiple Linear Regression	Multi Linear Regression
26.	Forecasting	Delhi	2009 to 2017	$PM_{2.5}$, PM_{10}, NO_2, CO, Ozone, and Benzene	Times Series Regression	Times Series Regression

Table1: Literature at Glance

1.3 Error Metrics

1.3.1 Mean Absolute Error (MAE)

An error is an absolute difference between the actual values and the values that are predicted. The absolute difference means that if the result has a negative sign, it is ignored. Hence, MAE = True values – Predicted values. MAE takes the average of this error from every sample in a dataset and gives the output. It is not very sensitive to outliers in comparison to MSE since it doesn't punish huge errors. It is usually used when the performance is measured on continuous variable data. It gives a linear value, which averages the weighted individual differences equally. The lower the value, the better is the model's performance.

1.3.2 Root Mean Square Error (RMSE)

RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model. In RMSE, the errors are squared before they are averaged. This implies that RMSE assigns a higher weight to larger errors. This indicates that RMSE is much more useful when large errors are present and they drastically affect the model's performance. It avoids taking the absolute value of the error and this trait is useful in many mathematical calculations. In this metric also, the lower the value, the better is the performance of the model.

1.4 Performance Metrics

1.4.1 R Squared

It is also known as the coefficient of determination. This metric indicates how well a model fits a given dataset. It indicates how close the regression line is to the actual data values. The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

2. Conclusion

Particulate Matter growth after a certain point may be anticipated using appropriate methodologies and precise data. In this study, Random Forest, Decision Tree, and a Regression model are the models for accurate prediction for their research, and Boosting technique was used to boost the model which grants power to machine learning models to improve their accuracy of prediction and most of the researchers were successfully established the model.

3. References

1. A. Masood, K. Ahmad, “A model for particulate matter PM2.5 prediction for Delhi based on machine learning approaches”, Elsevier, 2020.
2. B. Pan, “Application of XGBoost algorithm in hourly PM2.5 concentration prediction”, IEEE 2017.
3. C. Feng, W. Wang, Y. Tian, X. Que, and X. Gong, “Estimate Air Quality Based on Mobile Crowd Sensing and Big Data”, IEEE,2017.
4. D Petelin, A Grancharova, J Kocijan, “Evolving Gaussian process models for prediction of ozone concentration in the air”. Elsevier, 2013.
5. J. Ma, Z. Yu, Y. Qu, J. Xu, Y. Cao, “Application of the XGBoost machine learning method in PM2.5 prediction: a case study of Shanghai”, Aerosol and Air Quality Research. 2020.
6. J.K. Deters, R. Zalakeviciute, M. Gonzalez, Y. Rybarczyk, “Modeling PM2.5 Urban pollution using machine learning and selected meteorological parameters”, Journal of Electrical and Computer Engineering. 2017.
7. Jung, C Ren, B Hwang, and W Chen. “Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level PM2.5 concentrations in Taiwan from 2005 to 2015.” Elsevier, 2017.
8. K. B. Shaban, A. Kadri, and E. Rezk,” Urban Air Pollution Monitoring System with Forecasting Models”, IEEE,2016.
9. K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang, A. Rahman, “SVR Based Dense Air Pollution Estimation Model Using Static and Wireless Sensor Network”, IEEE, 2016.
10. KS Harishkumar, KM Yogesh, I. Gad “Forecasting air pollution particulate matter PM2.5 using machine learning regression model”, Elsevier, 2019.
11. Kumar A, Kamal, D, Maji, Jyoti, Deshpande, Ashok. “Disability-adjusted life years and economic cost assessment of the health effects related to $PM_{2.5}$ and PM10 pollution in Mumbai and Delhi, in India from 1991 to 2015.” Springer, 2017
12. Kumar S, Mishra S, Singh S, “A machine learning-based model to estimate $PM_{2.5}$concentration levels in Delhi's atmosphere”. Journal of Heliyon, 2020.
13. M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, S. Talebiesfandarani, “PM 2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data”, atmosphere, Multidisciplinary Digital Publishing Institute, 2019.
14. Nevin, G., Zur, G. “The regional prediction model of PM_{10} concentrations for Turkey” Journal of Physics: Conference Series, 2016.
15. O. Kisi, K. S. Parmar, K. Soni, and V. Demir, “Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, andM5 model tree models”, springer,2017

16. P. Wang, H. Zhang, Z. Qin, G. Zhang, "A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting", Atmospheric Pollution Research 2017.
17. Q. Di, H. Amini, L. Shi, I. Kloog, R. Silvern, J. Kelly, MB. Sabath, C. Choirat, P. Koutrakis, A. Lyapustin, Y. Wang, LJ. Mickley, J. Schwartz, "An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution", Environment International. ELSEVIER, 2019.
18. R. Waman Gore and D. S. Deshpande," An Approach for Classification of Health Risks Based on Air Quality Levels", IEEE, 2017.
19. Rubal, D Kumar, "Evolving Differential evolution method with random forest for prediction of Air Pollution". Elsevier, 2018.
20. S Yarragunta, M Nabi, Jeyanthi, Revathy, "Prediction of Air Pollutants Using Supervised Machine Learning". IEEE, 2021.
21. S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal, "Classification model for air quality using machine learning techniques," International Journal of Software Engineering and Its Applications, 2015.
22. Sarkawt M.L. Hama, Prashant Kumar, Roy M. Harrison, William J. Bloss, Mukesh Khare, Sumit Mishra, Anil Namdeo, Ranjeet Sokhih, Paul Goodman, Hemendra Sharma, "Four-year assessment of ambient particulate matter and trace gases in the Delhi-NCR region of India" ELSEVIER 2020.
23. Sharma N, Taneja S, Sagar V, Bhatt A, "Forecasting air pollution load in Delhi using data analysis tools". Journal of Procedia Computer Science, 2018.
24. T. W. Ayele, R. Mehta," Air pollution monitoring and prediction using IoT", IEEE, 2018.
25. W. You, Z. Zang, L. Zhang, Y. Li, X. Pan, W. Wang, "National-scale estimates of ground-level PM2.5 concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD" Multidisciplinary Digital Publishing Institute, 2016.
26. X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, J. Wang, "Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory-based geographic model and wavelet transformation". Elsevier, 2015.
27. Xi X, Wei Z, Xiaoguang, "A comprehensive evaluation of air pollution prediction improvement by a machine learning method". IEEE, 2016.
28. Y. Chen, "Prediction algorithm of PM2.5 mass concentration based on adaptive BP neural network", Springer, 2018.
29. Y. Rybarczyk and R. Zalakeviciute, "Machine Learning Approach to Forecasting Urban Pollution", IEEE, 2016.
30. Y.-F. Xing, Y.-H. Xu, M.-H. Shi, Y.-X. Lian, The impact of PM2.5 on the human respiratory system, Journal of Thoracic Disease, 2016.