Study on the Relationship Between Weather Components and PM 2.5 on Kolkata using Supervised Learning

Project report in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

In

COMPUTER SCIENCE & ENGINEERING

Submitted By

SOUMO BANERJEE	**University Roll No. 12018009002106**
PRADIPTA PAUL	**University Roll No. 12018009019500**
ANIMESH KUMAR SINGH	**University Roll No. 12018009019515**
SAHIN ALAM	**University Roll No. 12018009019164**
CHIRANMOY BHATTACHARYA	**University Roll No. 12018009019037**
RANJAN KUMAR SINGHA	**University Roll No. 12018009019523**

Under the guidance of

PROF. STOBAK DUTTA

Department of Computer Science and Technology

&

PROF. SUMIT KUMAR ANAND

Department of Computer Science & Engineering

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

CERTIFICATE

This is to certify that the project titled "**Study on the Relationship Between Weather Components and PM 2.5 on Kolkata using Supervised Learning**" submitted by Soumo Banerjee (University Roll No. 12018009002106, Pradipta Paul (University Roll No. 12018009019500), Animesh Kumar Singh University Roll No. 12018009019515), Sahin Alam (University Roll No. 12018009019164), Chiranmoy Bhattacharya (University Roll No. 12018009019037) and Ranjan Kumar Singha (University Roll No. 12018009019523) students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfilment of requirement for the degree of Bachelor of Computer Science and Engineering, is a bona fide work carried out by them under the supervision and guidance of Prof. Stobak Dutta & Prof. Sumit Anand during 8th Semester of academic session of 2021-2022. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original, and its performance is found to be quite satisfactory.

_____

Prof. Stobak Dutta

Assistant Professor

Department of Computer Science and Technology

UEM, Kolkata

_____

Prof. Sumit Kumar Anand

Assistant Professor

Department of Computer Science and Engineering UEM, Kolkata

_____

Prof. (Dr.) Sukalyan Goswami

HOD, Department of Computer Science and Engineering

UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. Stobak Dutta and Prof. Sumit Kumar Anand of the Department of Computer Science and Engineering, UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to Prof. (Dr.) Sukalyan Goswami, HOD, Computer Science and Engineering, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Soumo Banerjee

Pradipta Paul

Animesh Kumar Singh

Sahin Alam

Chiranmoy Bhattacharya

Ranjan Kumar Singha

TABLE OF CONTENTS

Study on the Relationship Between Weather Components and PM 2.5 on Kolkata using Supervised Learning

Soumo Banerjee, Stobak Dutta[#], Sumit Kumar Anand[#], Chiranmoy Bhattacharya, Animesh Kumar Singh, Sahin Alam, Pradipta Paul, Ranjan Kumar Singha

Department of Computer Science and Engineering & Department of Computer Science and Technology, University of Engineering and Management, Kolkata, India

Abstract: Air pollution has been a major threat to humanity and a matter of global concern for decades. In recent times Air quality of Kolkata has been termed hazardous many times. Climate change and air pollution is related very closely. The polluted Air has adverse effect on the human heath and ecosystem and also it is responsible for the climate change actively. Among the pollutants Particulate Matter is the most threatening to human health and it's amount in the Air of Kolkata is alarming most of the times in recent years. In this paper, we have studied the Meteorology and Air quality around the Victoria area (22°32'40"N 88°20'32"E) of Kolkata for the period of April 2018 to April 2022 and implemented various supervised learning models to study the relationship between Meteorological components and PM 2.5 in order to create a prediction model which may be used in remote parts of Bengal where AQI stations are not available to predict the PM 2.5 using only the meteorological data.

Keywords: Air Quality, PM 2.5; Supervised learning models; Artificial Neural Network; Machine Learning.

#Assistant Professor University of Engineering and Management

University Area, Plot No. III, B/5, New Town Rd, Action Area III, Newtown, Kolkata, West Bengal 700156

1 Introduction

Kolkata is very densely populated city (Census. 2011) in India and the capital of west Bengal. It is situated by the bank of river Ganga. For the Huge pollution and traffic different kind of pollutions tend to rise in Kolkata. In the last few decades, air pollution has become a matter of great concern for the city people all over the world (Chattopadhyay et al. 2010; Debone et al. 2020). The most effected cities in India due to AIR Pollution are Delhi and Kolkata (Dutta et al. 2020) It is found that 60% of Kolkata population is suffering from respiratory disease due to air pollution. The SO2 concentration is observed to be relatively low compared to other pollutants over Kolkata (WHO 1992). The air quality standard as annual average observed for 98% time of the year for NO2, SO2 and PM10, CO and O3 has been provided by Central Pollution Control Board (CPCB 2009). The major contributors to air pollution consist of mainly two types: gaseous air pollutants (SO2, NO2, CO, etc.) and suspended particulate matters (PM10, PM2.5, etc.). Among these various air pollutants, the deadly pollutant is PM2.5 (L. Miller at al. 2018). PM2.5 refers to the mass per cubic meter of air of particles with a size (diameter) generally less than 2.5 micrometers (μm) (Particulates. Wikipedia). PM2.5 is also known as fine particulate matter (2.5 micrometers is one 400th of a millimeters). Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high (Pandey et al. 2013). PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. In this study, the megacity Kolkata (22°32'40"N 88°20'32"E) is chosen for Air quality analysis. Meteorological and Air Quality data has been recorded for the period April 2018 to April 2022. Then after pre-processing several supervised learning models like Decision tree regressor, Random Forest regressor, Linear regression, Xgboost regression, Artificial Neural Network has been used to study the Nature of PM 2.5 with the Meteorological components and the models has been evaluated to predict the values of 2.5. The validity of the observation revealed that Random Forest is the best model to estimate the predictability of PM 2.5. However, little to no work has been done for predicting PM 2.5 using weather parameters for Kolkata Context.

2 Literature Review

It is inevitable to understand what has been done in the current field of work, to get an overall picture where the field currently stands. Many people till now have worked in studding the relationship between the Weather competes and the Air pollution and to predict a sustainable solution by applying soft computing methods.

The pollutant particles can have a severe impact on climate processes in terms of reflection, absorption, and scattering radiation, but among all the reflection and scattering are most important (Sarkar et al. 2006). The mixing or transportation of air pollutant throughout the atmospheric layers in the form of gas or solid particles generates the most fatal form of pollution (Kaushik et al. 2005). Various models have been implemented for the assessment of predictability, which are noteworthy. Jiang et al. in (2004) developed an ANN model to predict air pollution index from previous day meteorological parameters in replacement of old model for Shanghai city. The meteorological parameters have been found to be associated with the index (Cogliani 2001). The atmospheric pollutants and meteorological parameters have influences on the monsoon rainfall over some metropolises of India (Gunaseelan et al. 2014). Kurt et al. 2019 applied a deep learning approach to predict the leevel of harmful pollutants in air. They trained a simple neural network with a view to predicting air pollutants level ahead of three days. Their work is also similar to what we want to do in our paper, but using models like recurrent neural network. Their study area was greater Istanbul. They predicted accurate air pollution prediction using a simple neural network.. Bhalgat et al. 2008 applied machine learning algorithms to predict the amount of Sulphur dioxide. Basically, they used models in time series with a view to predict air pollutants in the air. They were also able to state which cities in India were highly polluted and which cities were less polluted. Their approach is very similar to ours as we are also applying machine learning algorithms for predicting the level of air pollutants but for Dhaka city context. Kaur et al. 2016 in their paper presented different big data and machine learning approaches that had been applied in air quality prediction till date. The authors summarized the latest achievements regarding air quality prediction. We were able to get a bigger picture of what is being done using machine learning and big data in air quality evaluation. Raj et al. 2018 presented the idea of using a feed forward neural network to see if it was able to predict the air pollution level. Our work is very similar to what they have done in their paper. They also used RNN for air pollution prediction. The novelty of their application was that they tried to predict the level of air pollution three days ahead.

3 Data and methodology

The study has been carried out with 4 years of data from April 2018 to April 2022 on the Meteorological data and Air Quality data of Victoria substation (22°32'40"N 88°20'32"E) of Kolkata. The Meteorological data has been collected from NASA weather Portal (https://power.larc.nasa.gov/data-access-viewer/) and the Air pollutant data (PM 2.5) data has been collected from the Central Pollution Control Board(https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing). In this paper we have used the daily average data of the aforesaid parameters. Here, we have considered Daily Average Temperature, Minimum and Maximum temperature, Precipitation, Relative humidity, Average Wind speed, Minimum and Maximum wind speed as our Meteorological parameters. On the other side, daily average of PM 2.5 is considered as the pollutant. The dataset is consisting of 8 independent features and 1 dependent feature (Fig: 1.1).

	T2M	T2M_MAX	T2M_MIN	PRECTOTCORR	RH2M	WS50M	WS50M_MAX	WS50M_MIN	PM2.5
9	28.29	34.65	23.40	5.95	67.94	5.19	7.70	2.67	34.49
10	29.12	36.17	23.28	2.08	66.56	5.68	7.59	3.39	27.89
11	29.19	36.48	23.39	7.67	67.56	5.38	7.50	3.14	21.64
12	29.26	36.38	23.40	5.68	66.50	5.20	7.64	3.12	23.07
13	30.48	38.27	24.51	2.63	63.50	5.32	7.64	3.66	21.27
...	...	...	...	...	...	...	...	...	...
1466	30.02	37.79	24.98	0.05	65.62	6.88	8.09	4.95	23.73
1467	30.11	38.51	23.83	0.20	65.12	6.16	7.30	5.41	26.85
1468	30.38	38.89	24.24	0.12	64.62	6.84	8.64	5.44	19.72
1469	30.98	39.46	25.00	0.32	64.88	6.47	8.62	3.92	11.15
1470	30.89	38.64	25.47	0.39	67.31	7.35	8.87	6.25	9.95

1395 rows × 9 columns

Fig: 1.1 (Dataset)

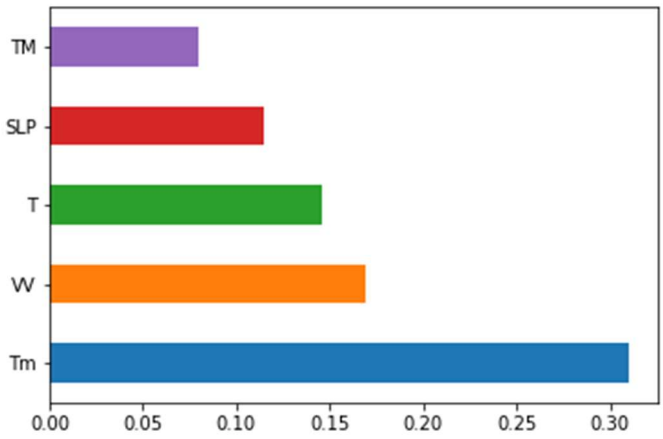

Fig: 1.2 Feature Importance

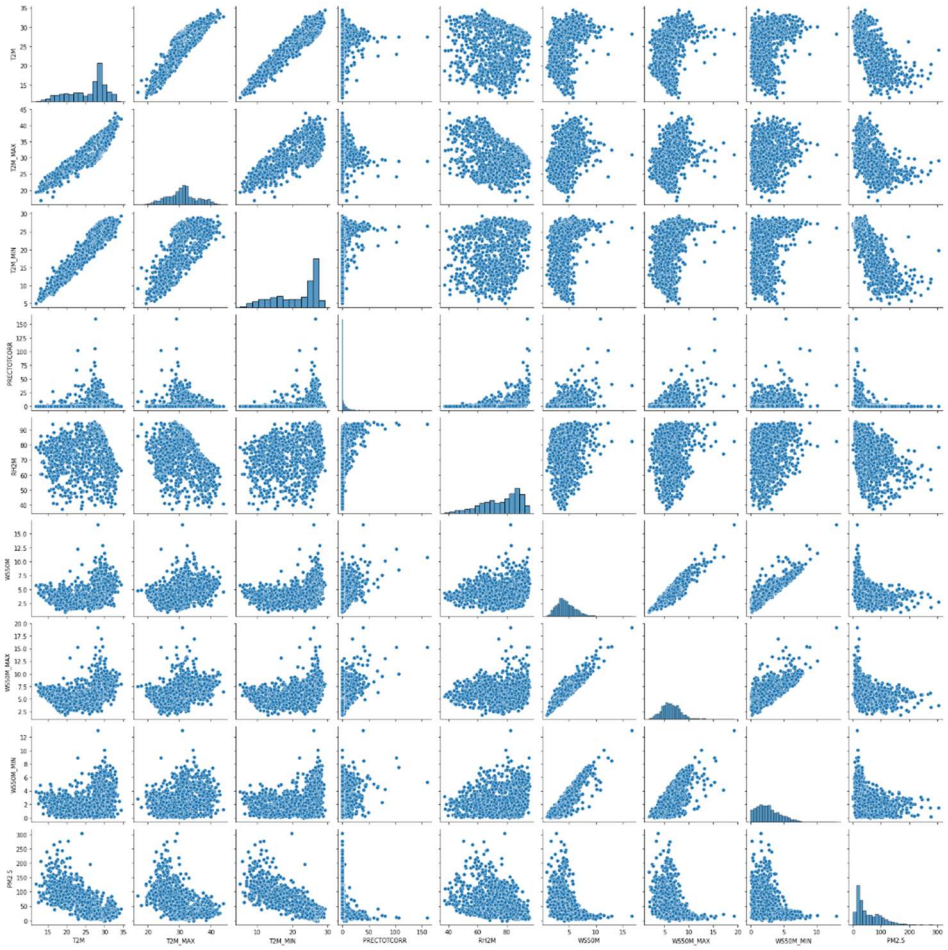

Fig: 1.3 Correlation Matrix

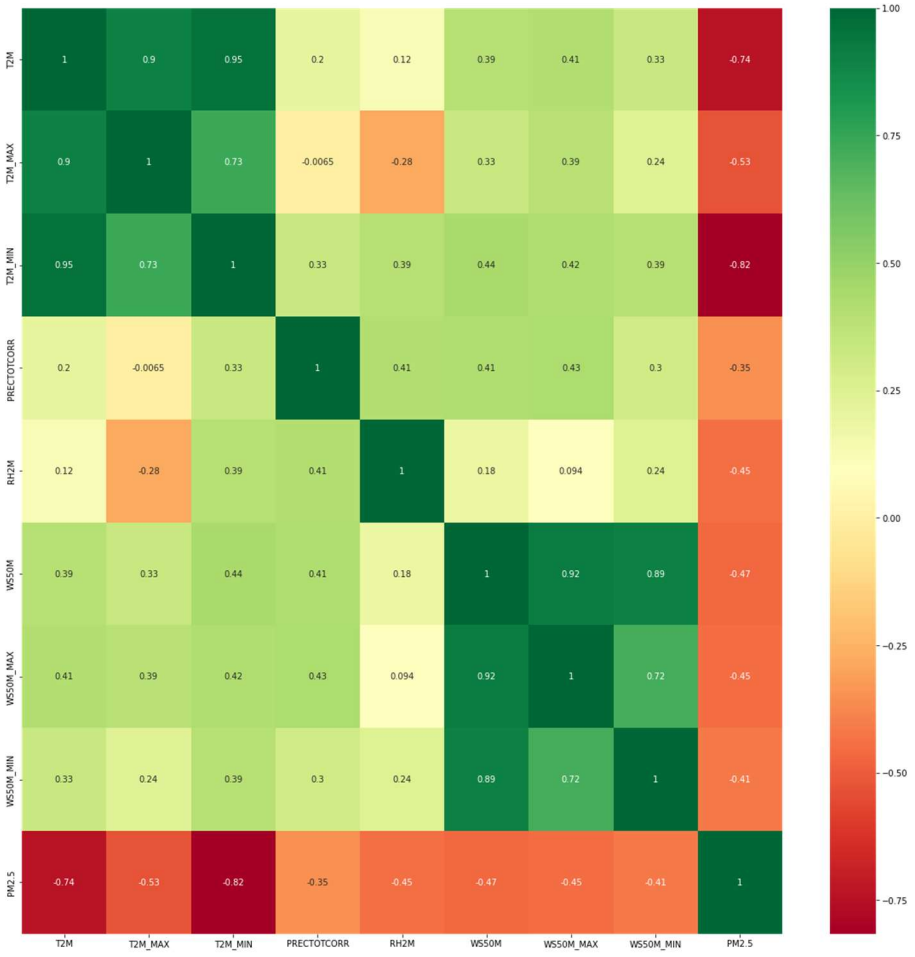

Fig: 1.3 Correlation Heatmap

The methodologies

Machine learning involves computational methods which learn from complex data to build various models for prediction, classification, and evaluation. The study attempts to build forecasting models capable of efficient pattern recognition and self-learning. In this section, the underlying principle of five machine learning methods as the canonical procedure will be discussed respectively.

3.1 Linear Regression

Linear regression is probably the method where most of the academicians started their first machine learning experience. Its main working principle lies behind the fitting of one or more independent variables with the dependent variable into a line in n dimensions. n usually denotes the number of variables within a dataset. This line is supposedly created as it would be minimizing the total errors when trying to fit all the instances into the line. Under machine learning, linear regression is equipped with the capability to learn continuously by optimizing the parameters in the model. These parameters are including w_0, w_1, w_2, ... , w_m (as illustrated in Figure 4). Most commonly, optimization is carried out by a method called gradient descent. It works by partially deriving the loss function and all parameters will be updated by subtracting the previous value with the derivative times a specified learning rate. The learning rate can be tuned by the simplest way, which is rule of thumb (trial and error), or a more sophisticated rule, e.g., meta-heuristic. Another parameter that is left for tuning is the amount of generalization added to the model. Regularization is undergone as an effort to lessen the chance of overfitting and increase the robustness of the model. Two types of regularization used in linear regression are lasso and ridge regression. Lasso regularization will eliminate less important feature by letting the feature's coefficient to zero and retain another more important one. Ridge regularization on the other hand will not try to eliminate a feature, but instead, tries to shrink the magnitude of coefficients to get a lower variance in the model.

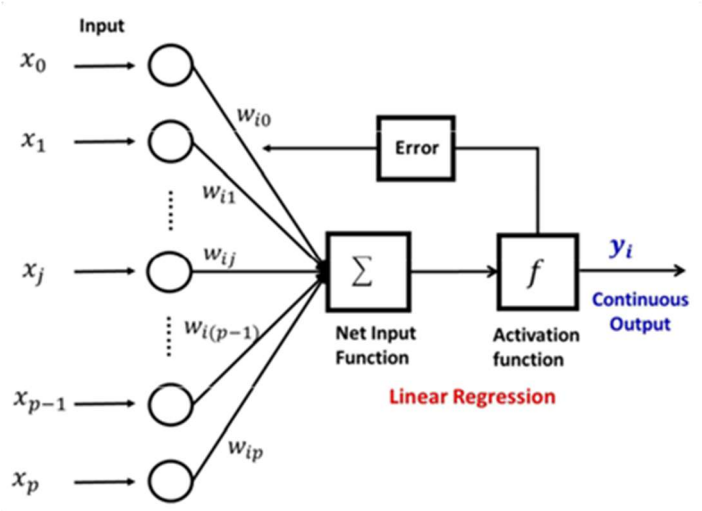

Figure 2.1. Demonstration of linear regression's learning process.

3.2 Decision Tress

The decision tree algorithm is a popular and simplest machine learning algorithm. It builds classification or regression models in the kind of a tree structure. The aim of the decision tree algorithm is to create a model that can be used to predict the class of the result variable by learning the decision rules inferred from prior training data. This algorithm uses if-then rule which is mutually exclusive and exhaustive for classification technique.

3.3 Random Forest

Another prominent machine learning method, random forest, a supervised learning ensemble algorithm, combines multiple decision trees to form a forest and the bagging concept, that latter adds the randomness into the model building. The random selection of features is used to split the individual tree while the random selection of instances is used to create training data subset for each decision tree. At each decision node in every tree, the variable from the random number of features is considered for the best split. If the target attribute is categorical, random forests will choose the most frequent as its prediction. On the other hand, if it's numerical, the average of all predictions will be chosen.

Random forest can tackle both classification and regression case. For prediction, each test data point is passed through every decision tree in the forest. The trees then vote on an outcome and the prediction is produced from a majority vote among the models and henceforth resulting in a stronger and more robust single learner. Random forests can overcome the prediction variance that each decision tree has, in the way that the prediction average will approximate

the ground truth (classification) or true value (regression). Figure 2 shows the illustration of a random forest that consists of m number of trees.

3.4 Adaptive Boosting

The next method, Adaptive Boosting, also came from a branch of ensemble methods where combine several weak learners yet with the sequential arrangement instead of a parallel setting as what random forest does. Boosting trains the base models in sequence one by one and assigns weights to the classifiers based on their accuracy to predict a random set of input instances. By such means, the more accurate classifiers will have more contribution in the final answer. The weights are also attributed to each input item depending on how difficult the instance to be predicted as on average by all classifiers. The higher the weight, the harder it is to estimate the ground truth for the instance and therefore this item will have a higher chance to appear as the training subset in the succeeding iteration. In other words, the boosting process concentrates on the training data that are hard to classify and over-represents them in the training set for the next iteration. The loop will start to be more substantial, as the focus is gathered to solve the difficult-topredict instances using the stronger classifiers. Classifiers are the base algorithms utilized to perform the prediction, where the common one used in AdaBoost is a decision tree. It also can be constructed from different types of algorithms, e.g., mix of a decision tree, logistic regression, and naïve Bayes (for classification).

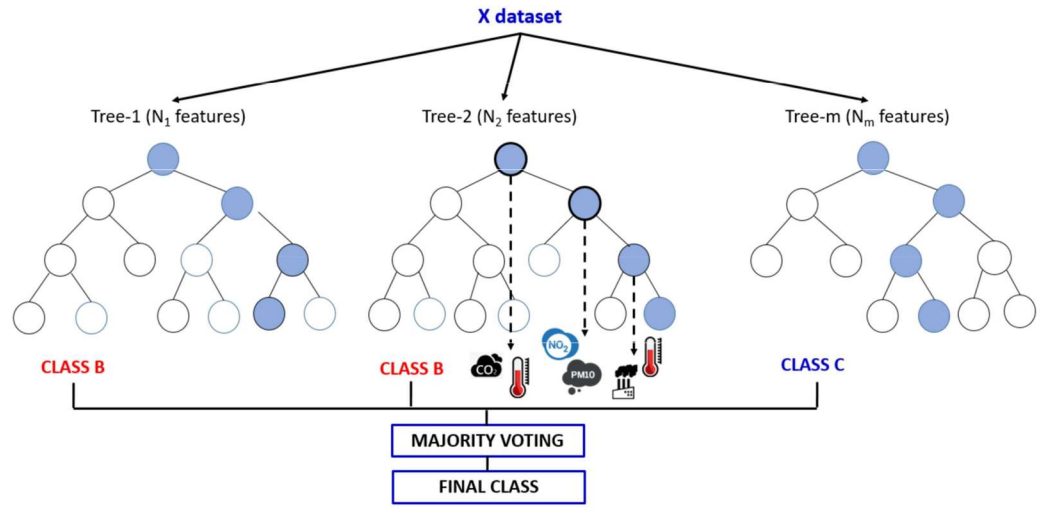

Figure 2.2 Illustration of a random forest algorithm.

3.5 Artificial Neural Network

The artificial neural work models have shown enormous potential in application field of various natures with the inclusion of back-propagation feed-forward learning technique (Chaudhuri 2015a, b). Artificial neurons or nodes are the basic processing elements of ANN. Connection weights represent each synapses and modulate the associated input signals where transfer function represents neurons/nodes which in turn exhibit the nonlinear characteristic of the event or system taken into account. Then, the neuron impulse is calculated as the weighted combination of the input signals and processed by the transfer function. A typical back propagating feed-forward neural network consists of input layer, hidden layers of hidden nodes, and an output layer. An output from a neuron is passed on to the next and thus for a given structured network is obtained (Chaudhuri et al. 2014a, b). ANN has been implemented worldwide real-world problems through pattern recognition, categorization, and forecast (Gardner and Dorling 1999; Chaudhuri 2006). Scientists have shown in their studies that ANN serves better than Box–Jenkins ARIMA model in forecasting the daily maximum ozone concentration (Yi and Prybutok 1996; Ghiassi et al. 2005). The ANN has also been implemented to forecast the track of tropical cyclones (Chaudhuri et al. 2014a, b), intensity of tropical cyclones (Chaudhuri et al. 2013), medium range forecast of cyclogenesis over North Indian Ocean (Chaudhuriet al. 2014a, b). The ANN models compared in this study are the multilayer perceptron (MLP), radial basis function network (RBFN), generalized regression neural network (GRNN), and linear neural network (LNN) (Fig. 2.4).

The most popular neural network model is MLP which is a feed-forward network normally trained using the back-propagation algorithm (Fig. 2a). This algorithm works on the theory of iterative process by changing connection weights, learning rate constant and by modulating number of layers and hidden nodes. This can be expressed as:

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \phi(w^T x + b),$$

where w is the vector of weights, x denotes the vector of inputs, b is the bias, and u is the activation function. RBFN is used as a network for its sensitivity and alternative to MLP (Hannan et al. 2010). This network consists of three layers: an input layer, a hidden layer, and an output layer (Fig. 2b). A Gaussian radial basis function is highly nonlinear and capable of

learning complex input–output mapping (Ma et al. 2007). The RBF for one hidden layer and a Gaussian RBF are represented as:

$$y_k(x) = \sum_{i=1}^{n} w_{ki} \exp\left(|- \frac{||x - u_i||^2}{\sigma_i^2}\right)$$

A GRNN model (Goulermas et al. 2007) has four layers: input layer, pattern layer, summation layer, and output layer (Fig. 2.c). The input layer is connected to the pattern layer where the training to produce output is processed. The pattern layer is connected to the summation layer. Both summation layer and output layer perform the normalization to produce output. This network is expressed as Gaussian function shown below (Hannan et al. 2010).

$$y_i(x) = \frac{\sum_{i=1}^{n} w_i \cdot \left[\exp - \left\{ \sum_{k=1}^{m} \left(\frac{x_i - x_{ik}}{\sigma} \right)^2 \right\}\right]}{\sum_{i=1}^{n} \exp - \left\{ \sum_{k=1}^{m} \left(\frac{x_i - x_{ik}}{\sigma} \right)^2 \right\}},$$

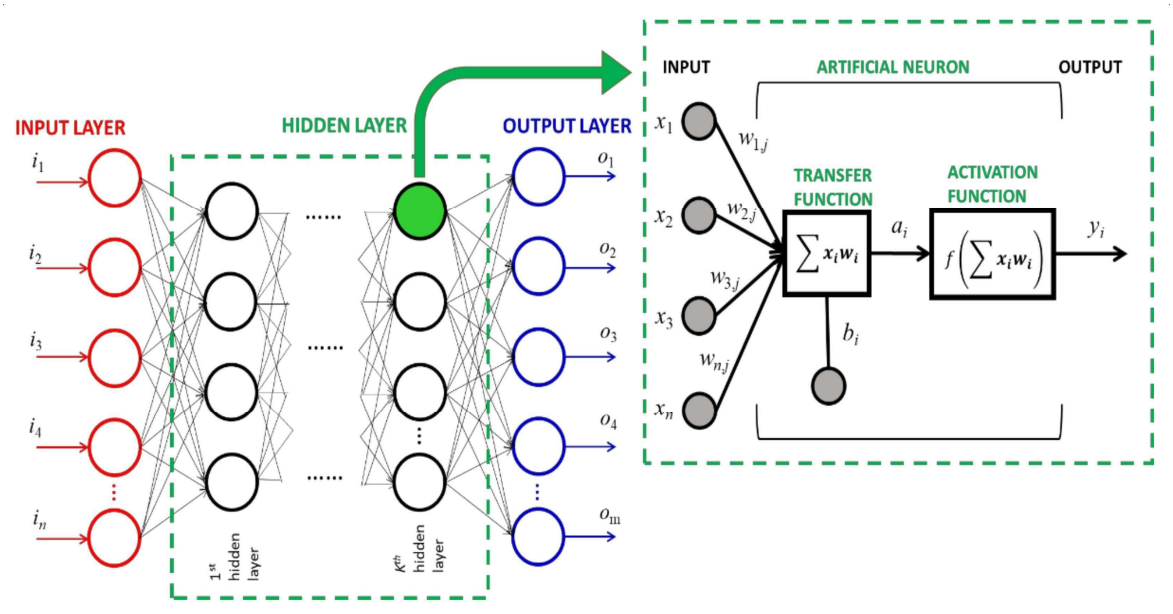

Fig: 2.3 Illustration of ANN

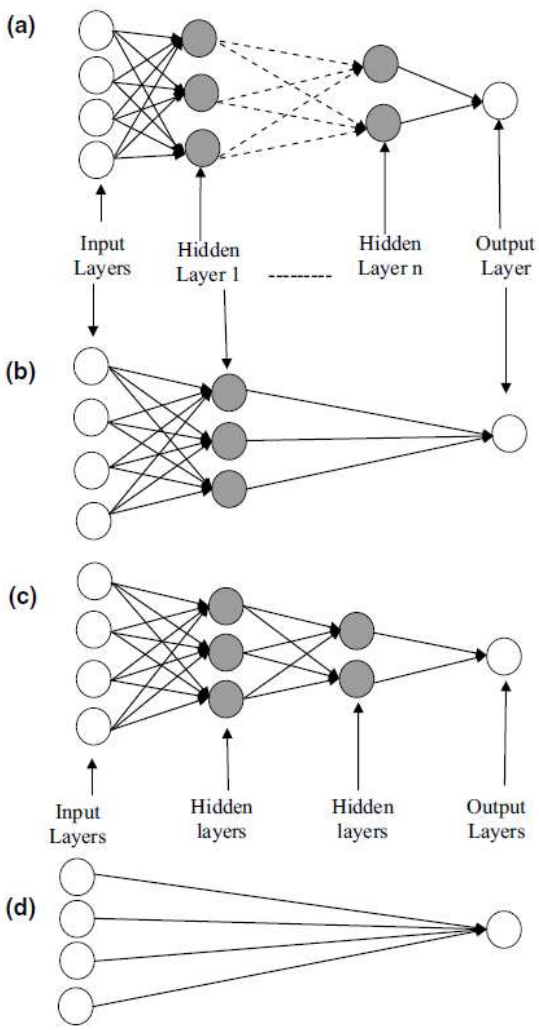

Figure 2.4 Structural representation of ANN models a. Multilayer Perceptron, b. Radial basis function network, c. Generalized regression nural network, and d. Linear

where wi is the weight connection between the neuron in the pattern layer and the summation neuron, n is the number of the training patterns, m is the number of elements of an input vector, xk and xik are the jth element of x and xi, respectively.

Linear neural network (LNN) is a neural network with two layers (Fig. 2d), input layer and output layer. It does not have any hidden layer or hidden nodes. LNN is represented by the following equations where the modification of connection weights is done (Yin et al. 2000).

$$w(k/k) = w(k/k-1) + K(k)d^T(k)$$
$$d(k) = y(k) - w^T(k/k-|1)b(k)$$
$$K(k) = P(k/k-1)b(k)[b^T(k)P(k/k-1)b(k) + r(k)]^{-1}$$
$$P(k/k) = [I - K(k)b^T(k)]P(k/k-1),$$

where $w(k/k)$ is connection weights, $K(k)$ is gain matrix, $d(k)$ is error vector, $P(k/k)$ is correlation matrix, and $r(k) = e - k/\mathrm{N}I$ (N is a constant).

3.6 Stacking Ensemble

Though coming from the same branch, stacking is quite different from the random forest and boosting strategy in AdaBoost in several ways. In bagging, variance in the final ensemble model is reduced by the random selection of a subset of features as well as instances for each predictor to execute the parallel and independent learning. The outcomes are then aggregated by the averaging method to generate an ensemble prediction. Boosting, on the other hand, will pass the dataset through all the learners which are set sequentially. Each instance and learner are given the attribute, the so-called weight, that is going to be updated on each pass (instance) and each iteration (learner). The weighting procedure results in the uneven contribution of each learner to the final prediction, and uneven prioritization to each instance for the training process - which substitutes the output averaging process mechanism and randomization for training

subset in the bagging concept. For stacking, each base predictor takes the whole dataset without any differentiation on the input and works in the canonical way to produce the result. The special property of this method lies in the aggregation mechanism. After the learning, the outputs from the predictors then become the inputs for the aggregator algorithm to produce the final prediction. The training set in the first learning process occupied by the base predictors is different with the one utilized by the aggregator algorithm because the dataset fed into the predictors has been transformed into the models which are later combined to form the new features. Fitting the aggregator algorithm onto the same instances causes a bias since the inputs are created from these instances. However, splitting two types of datasets raises another problem for a limited amount of data. To overcome this, the common k-folds cross-validation approach is usually adopted to provide more data for training both predictors and aggregators thereby facilitate a more accurate performance measure. In practice, stacking usually considers multiple types of learners to build the prediction, while bagging and boosting are more common to have only homogenous learners. Besides the algorithms used, the design of stacking ensemble can also be altered by the stacking level. If the number of levels is more than 2, the layer in the middle will be filled with multiple aggregators. However, since increasing the number of levels will cost on the time computation, this parameter usually remains in default (i.e., level size = 2).

3.7 K-Nearest Neighbor:

K-Nearest Neighbor is a simple algorithm that stores all the available data correspond to training data points in n-dimensional space and classifies the new data based on a similarity measure. Once an unknown discrete data is received, it analyzes the closest k number of instances saved and returns the most common class as the prediction and for real-valued data; it returns the mean of k nearest neighbors.

4. Experiments and Results:

Data pre-processing is used to convert the raw format of data into an understandable format because the data in the real world is incomplete, noisy, and inconsistent data. The generalized dataset undergoes pre-processing which helps to recover from missing values, null values, duplicate values, and convert the data into the numeric format. After data set is pre-processed the data set is divided into training and testing dataset. Then the dataset is trained using Linear Regression, Xgboost regressor, Lasso Regressor, Decision tree regressor, Random Forest, K

Nearest Neighbour regressor and Artificial Neural Networks. Then the accuracies are compared as below:

Algorithms	MAE	MSE	RMSE
ANN	18.11	780.34	27.93
Decision Tree	19.17	770.62	27.76
KNN	17.06	700.91	26.47
Lasso Regression	18.49	749.97	27.39
Linear Regression	18.64	754.22	27.46
Random Forest	16.81	663.93	25.77
Xgboost	16.99	689.14	26.25

The least maximum accuracy was achieved by Random Forest model after hyperparameter tuning.

Regression Accuracy Check (MAE, MSE, RMSE, R-Squared)

The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is that comparing the original target with the predicted one and applying metrics like MAE, MSE, RMSE, and R-Squared to explain the errors and predictive ability of the model.

1. Regression accuracy metrics
2. Preparing data
3. Metrics calculation by formula
4. Metrics calculation by sklearn.metrics

The MSE, MAE, RMSE, and R-Squared are mainly used metrics to evaluate the prediction error rates and model performance in regression analysis.

- Percentage Error

$$\mathrm{PE} = \frac{|(Yp - Ya)|}{Ya} 100$$

- MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.

$$\mathrm{MAE} = 1/n \sum_{i=1}^{n} |Yp - Ya|$$

- MSE (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$\mathrm{RMSE} = \sqrt{1/n \sum_{i=1}^{n} \overline{(Yp - Ya)}^2}$$

- R-squared (Coefficient of determination) represents the coefficient of how well, the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.

The air quality index or AQI (Kaushik et al. 2005; Rao 2014) was estimated using the formula used by Tiwari and Ali (1987). Initially the air quality rating (AQR) of each pollutant is computed:

$$Q = \frac{V}{Vs} \times 100,$$

where Q represents air quality rating for pollutants, V represents observed value of the pollutant, and Vs is the standard value recommended by national ambient air quality for every pollutant for the selected area. This rule suggests that if AQR is less than 100, then pollutant is within limit. If it exceeds 100, then the air of that area is polluted. The estimation of AQI is done by considering that if the total number of pollutants is 'n' for monitoring air quality, then the geometric mean of this 'n' number of pollutants is calculated in terms of their AQR:

$$g = anti\log \frac{(\log_a + \log_b + \log_c)}{n},$$

where '*g*' is the geometric mean of all pollutants used in the calculation and '*a,*' '*b,*' and '*c*' are the different AQR values of '*n*' number of pollutants.

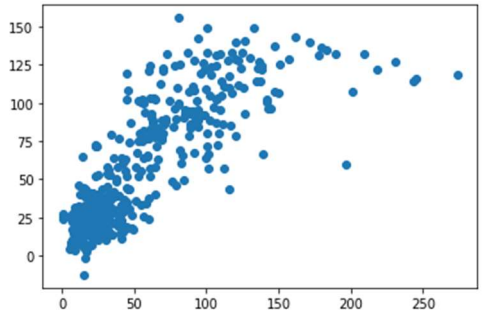

Fig: 4.1 Scatter Diagram of Linear Regression

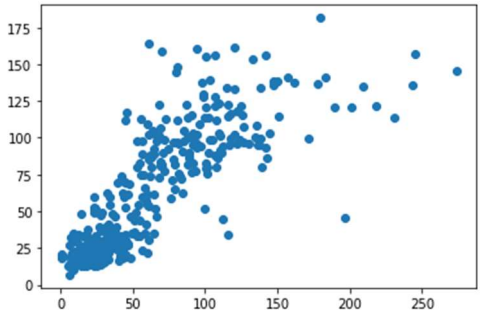

Fig: 4.2 Scatter Diagram of Xgboost

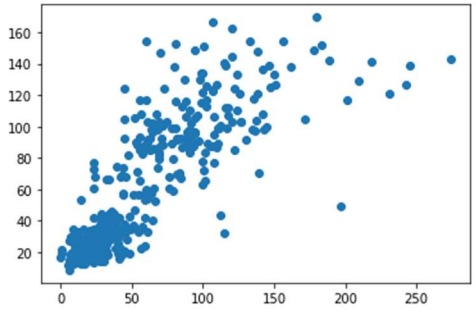

Fig: 4.3 Scatter Diagram of Random Forest

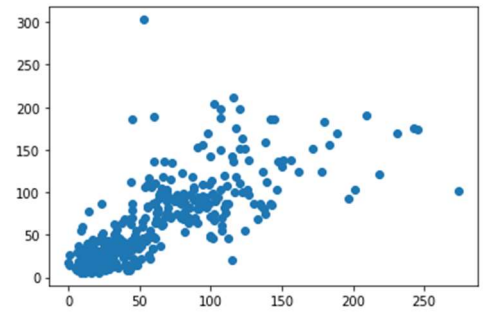

Fig: 4.4 Scatter Diagram of KNN

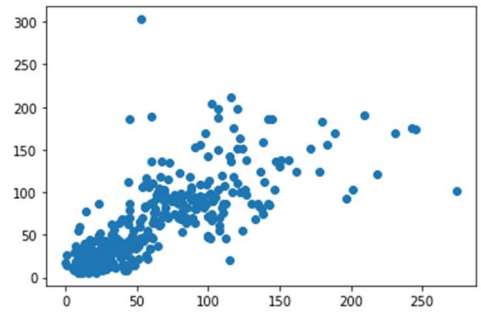

Fig: 4.5 Scatter Diagram of Decision Tree

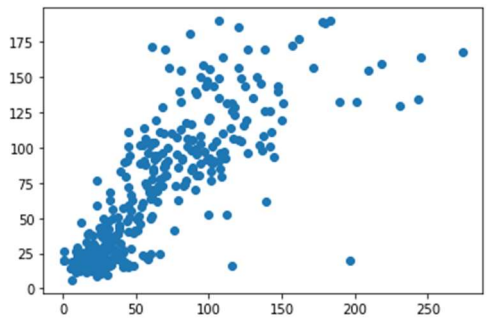

Fig: 4.6 Scatter Diagram of ANN

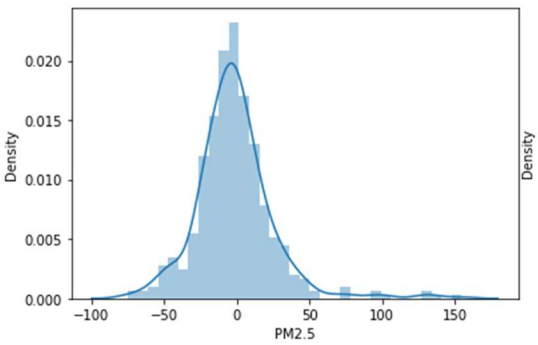

Fig: 4.7 Distribution graph after Linear Regression

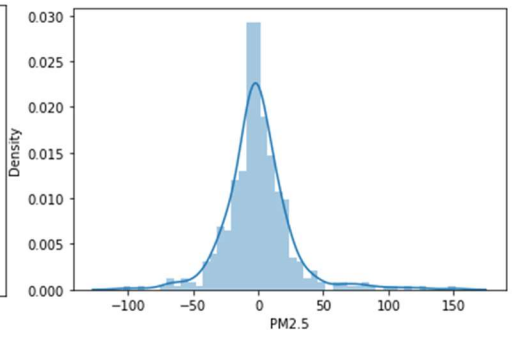

Fig: 4.8 Distribution graph after Xgboost

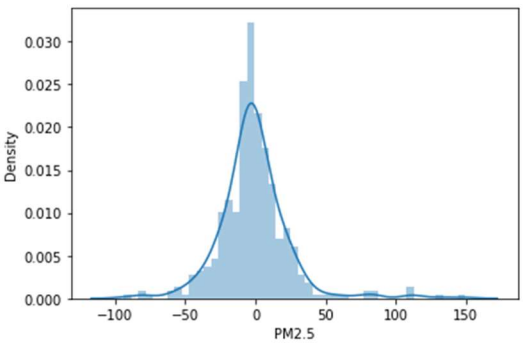

Fig: 4.9 Distribution graph after Random Forest

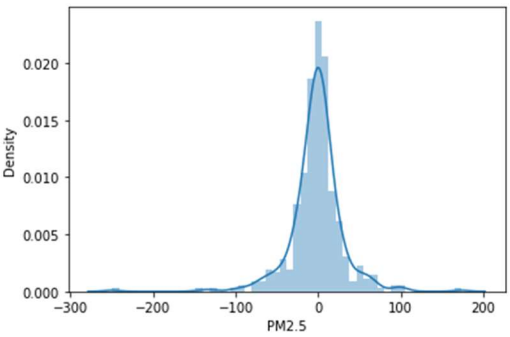

Fig: 4.10 Distribution graph after KNN

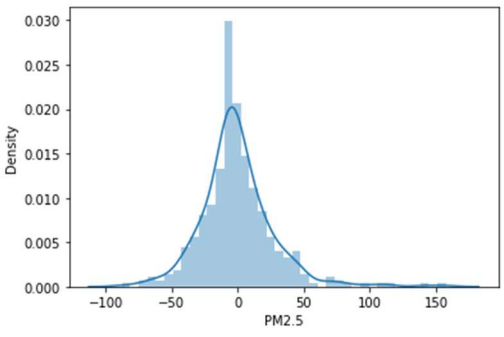

Fig: 4.11 Distribution graph after Decision Tree

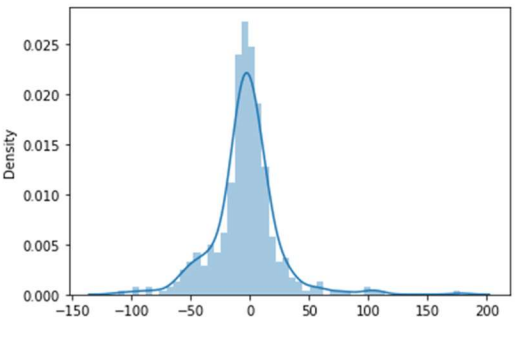

Fig: 4.12 Distribution graph after ANN

5. Conclusion

Air pollution causes many health effects on human beings as well as affects plants and animals. In this paper, the proposed system is developed to predict the PM 2.5 using supervised machine learning algorithms. The collected dataset is pre-processed to recover from missing, null, and duplicate values. The pre-processed dataset is divided into training and testing datasets in the ratio of 70:30 that is 70% of training and 30% testing dataset and in some instances 80:20. ML algorithms such as LR, Lasso Regression, K-NN, RF, and DT, Xgboost has been applied on the training dataset to train the dataset in order to obtain the highest accuracy. The performance measurement parameters like Mean Absolute Deviation (MAD), Root Mean Square Error (RMSE), Mean Squared Error (MSE) are calculated for each algorithm. The prediction model is constructed using the Random forest regressor. This prediction system helps asthma affected person to prevent themselves from the polluted area and also is developed to help the metrological department to predict air quality forecasting where the pollutant data is not available easily specially in the Rural areas of India. In the future, this air quality forecasting system can be optimized to implement in the Artificial Intelligence environment and can also automate this system by showing the prediction result in either web or desktop application.

References

Census (2011) Census of India 2011. Government of India. (http://censusindi a.gov.in/Digit alLib rary/Archi ve_home.aspx). Accessed 14 April 2022

Chattopadhyay S, Gupta S, Saha RN (2010) Spatial and temporal variation of urban air quality: a GIS approach. J Environ Prot 1(03):264

Shrabanti Dutta, Subrata Ghosh ORCID:orcid.org/0000-0001-5128-892X, SantanuDinda ORCID:orcid.org/0000-0003-3834-9841 *Aerosol Science and Engineering* volume 5, pages93–111 (2021)

L. Miller and X. Xu, "Ambient pm2.5 human health effects—findings inchina and research directions," *Atmosphere*, vol. 9, p. 424, 10 2018.

Particulates, https://en.wikipedia.org/wiki/Particulates

Pandey, Gaurav, Bin Zhang, and Le Jian. " Predicting sub-micron air pollution indicators: a machine learning approach.&quot ; Environmental Science: Processes & amp; Impacts 15.5 (2013): 996-1005

Kaushik CP, Ravindra K, Yadav K, Mehta S, Haritash AK (2005) Assessment of ambient air quality in urban centres of Haryana (India) in relation to different anthropogenic activities and health risks. Environ Monit Assess 122:27–40

Sarkar S, Chokngamwong R, Cervone G, Singh RP, Kafatos M (2006) Variability of aerosol optical depth and aerosol forcing over India. Adv Space Res 37:2153–2159

Jiang D, Zhang Y, Hu X, Zeng Y, Tan J, Shao D (2004) Progress in developing an ANN model for air pollution index forecast. Atmos Environ 38:7055–7064

Cogliani E (2001) Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. Atmos Environ 35:2871–2877

Chaudhuri, S., Chowdhury, A.R. Air quality index assessment prelude to mitigate environmental hazards. *Nat Hazards* **91,** 1–17 (2018). https://doi.org/10.1007/s11069-017-3080-3

A. Kurt, B. Gulbagci, F. Karaca, and O. Alagha, "An online air pollution forecasting system using neural networks," *Environment international*, vol. 34, pp. 592–8, 08 2008

P. Raj, "Prediction and optimization of air pollution-a review paper," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, pp. 3896–3904, 05 2019.

G. Kaur, J. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Bigdata and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, pp. 8–16, 01 2018.

T. Chiwewe and J. Ditsela, "Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations," 07 2016