Forecasting Air Quality for Different Air Pollutant Levels in Urban Areas Using Linear Regression Modeling

[1]Sayali Nemade, [2]Chandrashekhar Mankar

[2]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

Abstract: Air countenances our planet Earth to breathe in – it is the concoction of gases that augments to the air giving life expectancy to the globe and faunas that mark Earth such an ebullient domicile. Conversely, the gases that you are taking in day by day may perhaps be interminably slaying you. Bestowing to the World Health Organization, all over the place 2 million people give out from the impression of adulterated air every single year. Through the advancement of the economy and civilization all over the ecosphere, very nearly all cosmopolitan cities are coming into contact with augmented deliberations of ground-level air contaminants. Incited by the amassed sentience of the health facets in arrears to air effluence acquaintance exclusively by supreme profound consortiums of population such as progenies and the eldest inhabitants, short–range air pollution forecasts are delivered in inordinate extents by indigenous specialists. The Air Quality Index (AQI) is a numeral prearranged by constitutional bureaus to designate the eminence of the air at an assumed locality. AQI is used in lieu of resident and provincial air quality management in cosmopolitan cities entirely from corner to corner all around the world. Grounded on statistics from the preceding 6 years (2012-2018), the exploration spectacles that it relates to both the circadian limit of 50 micrograms per cubic meter and the tolerable number of per diem concentration fat chances to 35 per year. Correspondingly, the average almanac levels of PM10 overdid the approved customary of no more than 40 micrograms per cubic meter. For the reassurance of air contamination effluence, solid analyzing contexts are employed in copious borough areas. Making consumption of intelligence manoeuvres, in precise dominant frameworks and learning centred dealings can be believed as a boosting methodology. An evidence based attitude bids additional choice to the arithmetical simulations. In this paper, a linear regression model is proposed in which we take statistics from past 50 days and after analyzing it we are predicting intensities of diverse air pollutants 5 days in advance.

Index Terms **- Air Pollution, Air Quality Index (AQI), Linear Regression, Forecasting.**

I. INTRODUCTION

By this stretch, the ecological problem may possibly be the supreme solemn delinquent which has a pronounced influence on human well-beings and ecologies. The administrations have put forth marvellous determinations in the direction of the control of pollution, and have gained considerable triumph. For the reason that the use of gasoline and further petrochemicals and fossil fuels, air contaminants are evicted fundamentally by manufacturing subdivisions and automobiles. The realization of air impurities is an exact complicated and non-linear occurrence, owing to photochemical responses. Air pollution vitiates air superiority and centrals to severe maladies, such as asthma, bronchitis etc. Air Pollutant is shaped in the atmosphere when supplementary unswervingly emanated air contaminants conglomerate together. Even though Air Quality System monitoring data are comprehended as the ultimate standard for distinguishing ambient air quality and influential conformity by way of the government Air Quality Standards, principally this sort of data is inadequate in space and time. The extrapolation of the concentration of air impurities can educate the superior understanding of air effluence and arrange for useful statistics for the enlargement of most favourable emission control stratagems [1–5]. This prognostic knack would correspondingly provide a better understanding of the nature and relative comparisons of diverse emanation foundations that for all intents and purposes accountable for the ascertained levels of air pollutants. The arrangement which is competent to foretell the altitudes of air contaminants with ample prediction can arrange for individuals the time obligatory to cope the urgency. Massive progress has been made in the prophecy of the echelons of air pollutants over the preceding few epochs. Conversely, it is still a challenge to unerringly predict the levels of air contaminants due to the convoluted potent influences. It is indispensable to study additional proficient ways and means to approximately predict the levels of air pollutants in the nigh future.

Conspicuous levels of PM10 and PM2.5 can origin the augmented jeopardies of circulatory and respirational sicknesses [6]. The forecasting exemplary is prerequisite to make definite that these bounds are not surpassed, and if not, the evidence of the prediction will be of the essence for forthcoming environmental dogmas. The methodologies for the extrapolation of the intensities of different air contaminants can be disseminated into two categories: deterministic and stochastic. The deterministic methodology model the corporeal and biological transportation progression of the air pollutants in rapports of their efficacy of barometric variables such as wind rapidity, relative dampness, and temperatures with numerical models to foresee the levels of air impurities [7]. These methodologies can engender either short-run or long-run pollutant deliberation likelihoods. The functioning of these facsimiles depends upon comprehensive understanding of the formation course of pollutants. Some researchers strain to develop and mark superior a unified air quality modeling system that can prototype the sources, advancement, and ecological paraphernalia of different air pollutants by all scales. Nevertheless, it is still intriguing to unerringly predict the concentrations of diverse air contaminants for the motivation that the numerosity of sources and the complicated corporeal and biochemical processes affect the realization and shipment of these air contaminants. First of all the considerations in the equations have an imperative influence on the extrapolation enactment. In view of that, the intricacy of the hefty partial differential reckonings is quite high--they are arduous to solve exactly and will call for prodigious computation assets. In the time being, the concreteness and eminence of annotations which are used as inputs for the model also distress the guesstimate of numerical predictions. In this paper, we will predict PM2.5, PM10, NO_2, NH_3, SO_2 and CO by performing Linear Regression on them.

II. LITERATURE SURVEY

Arie Dipareza Syafei, Akimasa Fujiwara, and Junyi Zhang [8] In this paper, the panorama of each singular air pollutant as dependent variable was sought by employing slack 1(30 minutes beforehand), educated guess of air toxins (nitrogen dioxide, NO2, particulate matter 10um, PM10, and ozone, O3) and climatological elements and ephemeral dynamics as sovereign factors by considering successive error connections in the anticipated case. Selective dynamics improvement in consideration of self-determining segment examination and extensive segment investigation were resorted to procure detachments of the index factors to be chalked up into the model. Applying 30-mins improvised systemizations of NO2, PM10, and O3, they have shown the effects of different air toxins influence and barometric components. A comparative tactic in this paper can be combined by concentrating days inside week to the information from diverse stations in different borough areas to fortify prognostication.

Baltazar Frankovic, Viktor Oravec, Ivana Budinska [9] The paper focuses on an ontological arrangement for construction of air pollution control cognitive process base. The rudimentary reasons why the ontology methodology is recommended for validating learning base for air pollution control frameworks is that a symbolic cognitive process base can state the space controller's erudition without the peril that the comprehended information will be mislaid in an extraordinary measure of manageable reliable data. The metaphysics can be reclaimed in various comparable solicitation areas and for organization of other earthy issues.

Dr. S. W. A. Ashraf, S. Khanam , A. Ahmad[**10**] The previous investigation and interpretation making an allowance for the natural adulteration and its bearings on the human actualities, it can be well-grounded that the unreliable improvement of communal in conurbation together with over-crowding is first and foremost in charge of the miserable ecological state of affairs. Ensuing to previous scrutiny with respect to various indoor air impureness and their effects on wellbeing, it might be assumed that the indoor air impureness is affected by the housing conditions and living accommodations yet to some degree it is similarly affected by the outdoors circumstances.

Dan Wei [11] The control of air contaminant concentrations is hastily getting to be manifestly noteworthy amongst the most fundamental assignments for the statutory association of generating realms. This human activity strived to deploy some machine learning attitudes to guesstimate PM2.5 levels in veneration of a dataset encompassing every day typical weather and motility strictures in Beijing, China. The fundamental objective of the responsibility was the prognosis of air pollution intensities in Beijing City with rough approximation of informational index. The paramount machine learning method (SVM) yielded the 0.722 accuracy, 1.000 revaluation and 0.839 F-measure reverence.

I.N. Athanasiadis, K.D. Karatzas, P. A. Mitkas[12] Air superiority determination is one of the integral constituents of Urban Air Quality Management and Information Systems. The paper represents the interdependence work performed between a few mensurable approaches and order candidness, on the postulate of their performance for peculiar air quality time arrangement in Athens, Greece. The close exploration of the models performance presented that for the specific experimentation the consortium calculations have an extendedly better performance compared with the existent methods. Attempts yet to come will single-mindedly focus on conjoining unambiguous arrangement models into accrued classifiers, (i.e. gathering cognitive process).

Justin R. Chimka, Ege Ozdemir. [13] A linear regression model of particle contamination and an entreated logistic regression model of the indispensable top score were taken into contemplation for exemplifications in the US city of Los Angeles, California. Models were realistic to stature Air Quality Index (AQI) from an illustration, and thorough exploration was performed on them. Linear regression facsimiles of AQI through constituent part contamination are ultimately endowed to antedate target air quality; called on behalf of logistic regression models of AQI unambiguously are much likely supported to guesstimate astonishing air quality.

Mihaela M. Oprea[14] The paper represents an ontology for air pollution scrutiny and control, air pollution Onto, and describes its usage in two contextual probes, a specialist framework, and a multi-agent framework, both dedicated to inspection and control of air pollution in borough areas. The usage of the ontology in a specialist framework renders how protracted the information is to be based that will fundamentally be sensible, unambiguous and comprehensive although in case of a MAS it is a dogmatic assistance for amongst operators correspondence.

Ofoegbu E.O.,Fayemiwo M.A, Omisore M.O[15] To ramp up an air impureness influential solicitation framework for scrutinizing and determining air contaminants statistics with an explicit end objective to give information about the indispensable eminence of air we breathe in. Air effluence is the demonstration of compounds, constituent part, carbon-based constituents, or other precarious constituents into the earth's surface (Wikipedia, 2001). Nigeria is perceived with such a generous amount of varieties of air pollution, even though the crucial distress of this exploration effort is air pollution triggered by up-to-the-minute emancipations. The AQMS solicitation indoctrination exhausts the investigation of air contaminants information to reckon air quality and for predicting statistics, it allows personages in a definite constituency to screen the superiority of air they breathe in. This paper can correspondingly envoy supremacy to the enhancement of additional air discerning framework in diverse federations/realms to be generated which ought to be consummated through any sort of programming paradigm.

III. REGRESSION EXPLORATION

Regression analysis is an ongoing module for the exploration of associations between components. Mostly over and over again, the researcher attempts to treasure trove the perfunctory influence of one variable upon another variable—the upshot of a cost increment upon request, for example, or the bearing of changes in the cash supply upon the elaboration rate. To scrutinize such issues, the mediator pulls together statistics on the ultimate components of insurance premium and make usage of retrogression approach to evaluate the reckonable impact of the unpremeditated components on the variable that they touch on. The expert furthermore gauge the “existing enormity” of the evaluated acquaintances, that is, the level of surety that the authentic liaison is for all intents and purposes close to the evaluated affiliation. Stepwise linear regression is an attitude for reverting different components usually over an extended period kicking out those that aren't essential. Forward stepwise regression incorporates commencement deprived of any components in the model, examining the choice of every single component ultimately using a selected show scrutiny prototype, which includes the variable (say any) that potentiates the model furthermost, and reclaiming this technique up to the time that none intensifies the model.

3.1 Essential Postulates & Properties of Regression

Equally perceived, the utilisation of the base SSE extent might be protected on two grounds: its computational state of affairs and its fascinating ongoing properties. We might ruminate that these properties and the assumptions that are imperative to pledge them [16].

The notion is that income in "this present circumstances" are settled irrefutably by means of the condition I = α+ ßE + X +ε —spot-on values of α, ß, and ε occur, and we want to determine the existence of what they are. Making an allowance for the noise term e, whether that is spot-on or not, we can solely stature these considerations. We can take into contemplation the noise term e as a randomized variable, essentially derived from some likeliness circulation—individuals gain an instruction and collect together work apprehension, at that point nature creates a random number for every individual, called e, which increments or deducts salary as per the needs. When we take into contemplation the noise term as an irregular variable, it turns up to be definite that the assessments of α, ß, ε and (as accredited by their presently prevailing reverences) will similarly be random factors, in respect of the fact that the appraisals given by the SSE model will be contingent upon the peculiar approximations of e derived from nature for every single individual in the informational assemblage. Similarly, in view of the datum that there subsists an analogous circulation from which every single e is derived, there must correspondingly occur a likeliness approximation from which each consideration assessment is strained, the latest distribution of an element of the prior dispersals. The pleading presently existing properties of revert all distress the connection between the similar circulation of the parameter competences and the authentic guesstimates of those constraints. We requisite to flinch with a few delineations. The base SSE model is entitled as an estimator. Selection basis for producing parameter gauges, (for example, bounding all mistakes in ultimate regards) are similar estimators. Every single parameter competence that an estimator gives equally can be beheld as a random variable derived from some similar circulation. On the condition of not being likely that the mean of that similar dispersions is in actual fact equal to the authenticate estimates of the parameter that we are attempting to assess, at that point the estimator is unprejudiced. Per se, to revert back to our expositions, foresee making a continuation of informational assemblage each containing analogous individuals with analogous approximations of training and experience, differing precisely in that nature draws an alternate e for every individual for every single dataset. Anticipation progress that we re-evaluate our parameter competences for every single data set, to such a notch creating an outlook of assessments for each and every consideration. In a long shot, the estimator is unprejudiced, we would largely remunerate the authentic approximations of each and every parameter. An estimator is well thought-out as reliable in hundred-to-one-shot if it exhausts additional information to pronounce more apposite estimations. Ensuing the statement unambiguously, a sturdy estimator vintages assessments that converges upon the authentic estimates of the hidden parameter as the specimen measure gets relatively greater. Accordingly, the similar approximation of the competence for any consideration has put forth variation as the specific instance measure increments, and furthermost (inestimable test guesstimate) the capabilities will upsurge to the authentic honours. The metamorphosis of an estimator for a given specific occurrence measure is ultimately enthralling. Predominantly, let us limit to the intent consideration concerning estimators that are unprejudiced. Furthermore, take down variations in the similar dispersions of the estimator is noticeably tantalizing — it decrements the likeliness of a stature that direct contrast exceptionally from the authentic guesstimate of the rudimentary parameter. In visual perception at modified non-discriminatory estimators, the one with the most petite variation is entitled prolific or best.

IV. LINEAR REGRESSION

Linear Regression is a machine learning methodology grounded on supervised learning. It implements a regression job. Regression simulates an objective prediction value that is grounded on independent variables. It is fundamentally used for finding out the affiliation between variables and prediction. Different regression models swerves based on – the kind of rapport between dependent and independent variables that they are making an allowance for and the numeral of independent variables that are being used. Linear regression accomplishes the job of envisaging a dependent variable value (y) constructed on a given independent variable (x). So, regression routine treasure trove a linear rapport between x (input) and y (output). Henceforward, the appellation is Linear Regression. A few rudimentary perceptions of indicators before we flinch by way of linear regression:

1) Correlation (r) – Correlation elucidates the affiliation between two variables, conceivable standards are -1 to +1
2) Variance (σ^2) – Variance is the extent of spread in your statistics
3) Standard Deviation (σ) – Standard deviation is the quantity of spread in your statistics (Square root of Variance)
4) Normal distribution – The residual obligates to be customarily distributed _
5) Residual (error term) – Assessment of residual is equivalent to {Actual value – Predicted value}

4.1 Expectations of Linear regression: Not the identical magnitude fits for all, the equivalent applies to Linear Regression as well. In mandate to be kowtowed, a linear regression line statistics should sentient to few rudimentary but then again significant expectations. If your statistics doesn't coincide with these conventions then your domino effect may be erroneous as well as point in the wrong direction.

1) **Linearity & Additive**: There must be a rectilinear affiliation between dependent and independent variables and the influence of change in independent variable values should partake undeviating impression on dependent variable.
2) Normality of error distribution: Scattering of metamorphoses between actual and predicted values ought to be routinely dispersed.
3) Homoscedasticity: Variance of errors necessitates to be constant versus,
a) Time b) the predictions and *c) the independent variable values.*
4) Statistical independence of errors: The inaccuracy terms i.e. the residuals ought not to have any correspondence in the midst of themselves.

4.2 Linear Regression Line*:* While undertaking linear regression our end goal is to apt a line from end to end with the dissemination which is flanking to greatest of the points. Henceforward plummeting the aloofness i.e. the miscalculation term of statistics points from the fitted line. In lieu of case in point, in figure 1 specks exemplify innumerable data points and line epitomizes a contiguous line which can elucidate the affiliation between 'x' & 'y' axes. Exploiting linear regression we attempt to treasure trove such a formfitting line. For case in point, if we devour one dependent variable say 'Y' and one independent variable say 'X' then the rapport between 'X' & 'Y' can be embodied in a form of equation:

$Y = B_0 + B_1X$

Where Y is Dependent Variable, X is Independent Variable, B_0 is Persistent term i.e. Intercept and B_1 is Coefficient of correlation between 'X' & 'Y'.

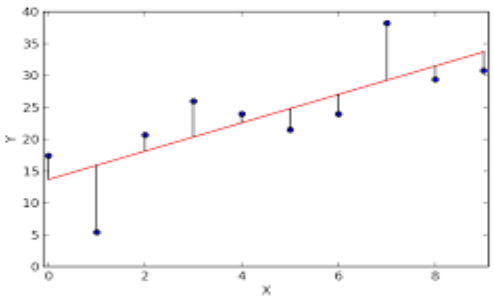


Figure 1 Linear regression Line

4.3 Characteristics of Linear regression Line*:*

1) Regression line have a duty to always pass from end to end the mean of independent variable (X) as well as mean of dependent variable (Y).

2) Regression line underestimates the entirety of "Square of Residuals" aimed at this intention the technique of Linear Regression is correspondingly acknowledged as "Ordinary Least Square (OLS)"

3) B_1 elucidates the amendment in Y condition the value of X vicissitudes by one entity.

4.4 Decree on the Linear Regression Line*:* Expending an arithmetical contrivance say for instance Excel, R, SAS etc. we can exactly so treasure trove the values of constants B_0 and B_1 as an end result of linear regression function. Let's say we want to envisage 'Y' from 'X' in the succeeding table 1 and for that our linear regression utility will look similar to "$y= B_{0+} B_1x$"

Table 1: Forecasted Y from X

x	y	Forecasted 'y'
1	2	$B_0 + B_1 *1$
2	1	$B_0 + B_1 *2$
3	3	$B_0 + B_1 * 3$
4	6	$B_0 + B_1 *4$
5	9	$B_0 + B_1 *5$
6	11	$B_0 + B_1 *6$
7	13	$B_0 + B_1 *7$
8	15	$B_0 + B_1 *8$
9	17	$B_0 + B_1 *9$
10	20	$B_0 + B_1 * 10$

Where,

Table 2: Standard Metrics for X and Y

Standard Deviation for x	3.02765
Standard Deviation for y	6.617317
Mean value of x	5.5
Mean value of y	9.7
Correspondence between x and y	0.989938

If we differentiate the Residual Sum of Square (RSS) with reverence to B_0 & B_1 and parallel the outcomes to zippo, we will get the ensuing equivalences as an outcome:

*B_1 equates to Correlation * (Std. Dev. of y/ Std. Dev. of x) and*

*B_0 equates to Mean(Y) – B_1 * Mean(X)*

Substituting values from table 2 and placing it into the directly above reckonings we acquire,

B_1 equivalent to 2.64, B_0 equivalent to -2.2

Hence, the slightest regression reckoning will look like

Y *equates to* **-2.2 + 2.64*X**

Here and now let's see how our extrapolations will look like expending this equivalence:

Table 3: Actual and Predicted values of Y

X	Actual Y	Predicted Y
1	2	0.44
2	1	3.08
3	3	5.72
4	6	8.36
5	9	11
6	11	13.64
7	13	16.28
8	15	18.92
9	17	21.56
10	20	24.2

Now we have merely 10 data points to apt a line as a result our extrapolations are not pretty accurate but then again if we comprehend the correspondence between 'Actual Y' & 'Predicted Y' here it will turn up to be very high; hence both the successions are moving unruffled and the graph for envisioning our anticipated values as displayed in figure 2:

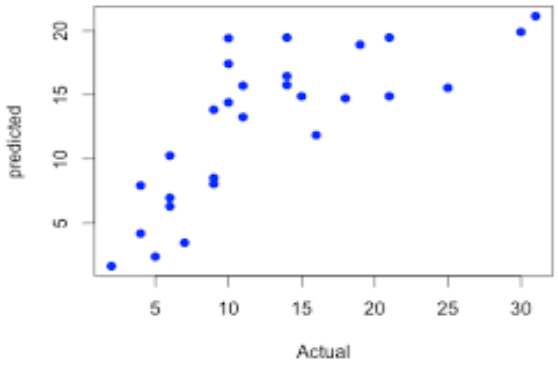


Figure 2 Graph of Actual Y and Predicted Y

4.5 Prototypical Enactment: The minute we structure the model, the succeeding stage is to discern whether your exemplary is precise enough to envisage in future or the affiliation which we ingrained in the middle of dependent and independent variables is conventional or not. Intended for this purpose there are several metrics which we will gaze into:

1) **R2:** Formulary for conniving R^2 is specified by

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

2) **Total Sum of Squares (TSS):** TSS is an extent of entire variance in the dependent variable Y and it can be believed as the expanse of variability that is ingrained in the response afore the regression is accomplished.

3) **Residual Sum of Squares (RSS):** RSS measures the expanse of unpredictability that is left inexplicable when we implement the regression. (TSS – RSS) measures the extent of changeability in the response that is elucidated when we execute the regression where N is the number of interpretations used to conform to the exemplary and σ_x is the customary deviation of x and σ_y is the customary deviation of y. A small number of opinions to keep in mind:

a) R^2 variates from 0 to 1.

b) R^2 of 0 revenues that the dependent variable cannot be anticipated from the independent variable

c) R^2 of 1 revenues the dependent variable can be foretold from the independent variable deprived of an error and

d) R^2 amongst 0 and 1 designates the bound to which the dependent variable can be projected. An R^2 of 0.20 means that 20 percentage of the variance in Y can be anticipated from X.

4) **Root Mean Square Error (RMSE):** RMSE articulates the extent of dissemination of forecasted values as of actual values. The procedure for computing RMSE is

$$R^2 = \{ (1/N)* \sum [(x_i - \text{mean}(x))* (y_i - \text{mean}(y))]/ (\sigma_x/ \sigma_y) \}^2$$

Where N is Overall number of annotations

Although RMSE is a virtuous extent for miscalculations but the concern with it is that it is susceptible to to the assortment of dependent variables. If dependent variable has diminutive range RMSE will be truncated and if dependent variable has comprehensive range RMSE will be extraordinary. Hence, RMSE is a decent metric to compare amongst different reverberations of an exemplary.

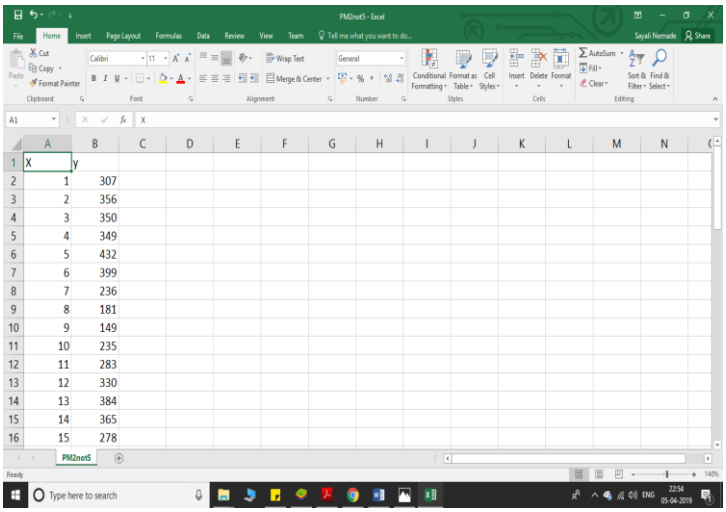
5) **Mean Absolute Percentage Error (MAPE):** In mandate to overcome the precincts of RMSE, just about all predictors have a preference for MAPE over RMSE which provides error in rapports of percentages and henceforward it is comparable across facsimiles. Formula for computing MAPE is

$$RMSE = \sqrt{\frac{\sum (Y_{Actual} - Y_{Predicted})^2}{N}}$$

V. RESEARCH METHODOLOGY

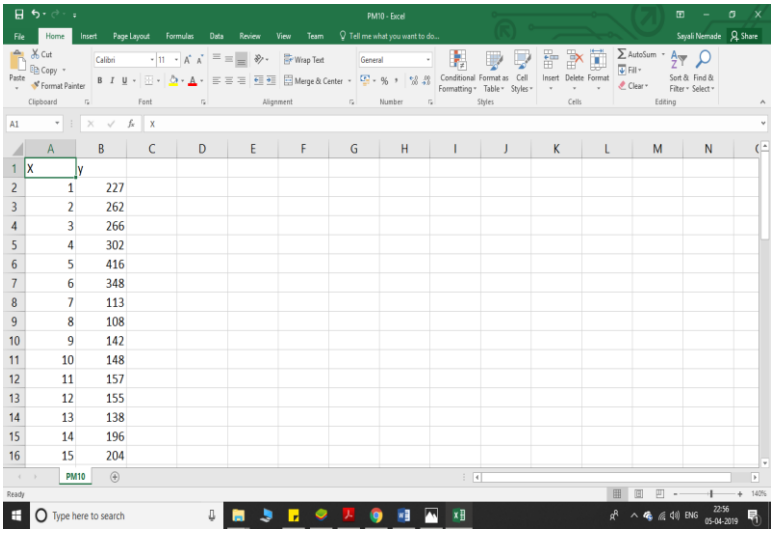
The data for this project is being collected from the site: **https://app.cpcbccr.com/AQI_India/.** We collected data for 50 days from 1st of February 2019 till 2nd of April 2019 that included concentrations of different air pollutants i.e. PM2.5, PM10, NO_2, NH_3, SO_2 and CO along with their average values. Based on this dataset we performed linear regression on each air pollutant and predicted their values for 5 days in advance. The data collected for this project was for Anand Vihar, Delhi DPCC around 12:00pm. So each air pollutant had like 61 data points including NaN values.

We created different files for each air pollutant so that we can get the predicted values for each of the air toxins. Datasheet for PM2.5 concentrations is shown in figure 3 which contains two rows x and y. x represents the value for 'day' and y represents the value for 'average value of PM2.5 concentrations' for that specific day around 12:00 pm. Similarly, the separate datasheet for PM10 concentrations is shown in figure4 given below. We created the same datasheets for remaining air pollutants i.e. NO2, NH3, SO2 and CO. We had to perform data cleaning before performing any analysis on the dataset. The coding was done in python for which the IDE used was Visual Studio Community 2017. The python environment that we implemented also included Anaconda 5.2.0 for executing our programs.



X	y
1	307
2	356
3	350
4	349
5	432
6	399
7	236
8	181
9	149
10	235
11	283
12	330
13	384
14	365
15	278

Figure 3 Datasheet for PM2.5 concentrations



X	y
1	227
2	262
3	266
4	302
5	416
6	348
7	113
8	108
9	142
10	148
11	157
12	155
13	138
14	196
15	204

Figure 4 Datasheet for PM10 concentrations

We accomplish the following steps here:

5.1 Loading the Data and Importing Libraries*:* Foremost we needed to load the data into the data frame and import pertinent libraries. The important libraries were pip, numpy, scipy, pandas, scikit-learn, matplotlib and sklearn. These were the required header files for the coding.

5.2 Data Cleansing: Most of the stints, in real dataset there will be outliers, erroneous data and even typing errors. Real dataset can also take account of NaN fields, data may not be sufficient for calculating the AQI or data may not be available for a particular day or days at all. So data cleansing is the fundamental part before scrutinizing your data. We needed to remove unsolicited fields as we are envisaging air contaminant deliberations 5 days in advance. So we need to look at what data we are actually looking at while constructing extrapolations. Removing blank values i.e. NaN's in Panda data frames is quite imperative as numerous machine learning procedures can't use these as inputs. Our data frame had quite a number of data fields with Nan's. After removing the NaN fields, rows with 'No data available' and certain rows for which the data was insufficient to calculate the AQI our dataset was considerably smaller i.e. each air pollutant had 50 data points each. Now we had a clean dataset with no misplaced values which can be analysed with the linear regression function.

5.3 Splitting Dataset into Training and Test Data: Cross validation is at all times prerequisite when training machine learning methodologies to be able to trust the platitude of the model generated. We will for all intents and purposes split our data into training and test data expending

scikit-learn's built in apparatuses. Also in lieu of scikit learn we need to detach our dataset into inputs and the feature that are being anticipated i.e. X's and Y's.

V. RESULTS AND DISCUSSION

The main air pollutants for Anand Vihar, Delhi DPCC were PM 2.5, PM 10, NO2, NH3, SO2 and CO. For some days PM 2.5 was prominent pollutant and for some days PM 10 was prominent pollutant. CO was also the prominent air pollutant for few days. However, the levels of remaining air toxins were also high to moderate for quite a large number of days. Let's see the results of our implemented model.
1) Analysis report of PM 2.5 is shown in table 4:-

Table 4: Analysis Report of PM2.5

Intercept	306.961616
Slope	[-2.78434343]
Mean Absolute Error	57.64055
Mean Squared Error	5184.57095
Root Mean Squared Error	72.003964
R^2, Co-efficient of Determination	0.273688
R	0.5231523

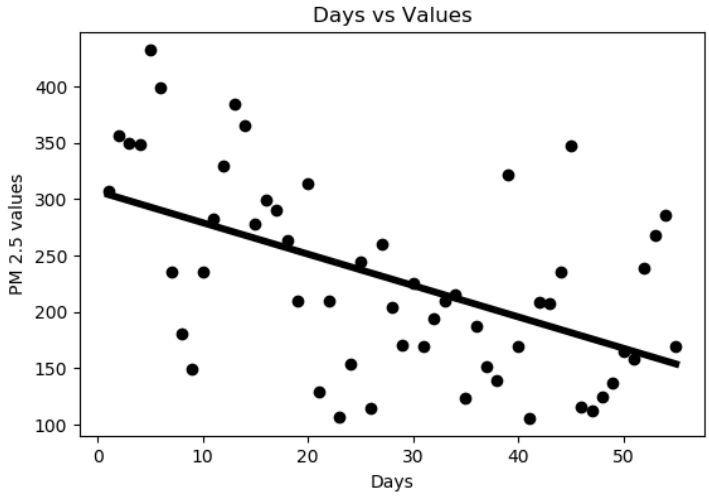


Figure 5 Graph for Days vs PM 2.5 Values

Forecasted values for PM 2.5 for next five days are shown in table 5:

Table 5: Forecasted values of PM2.5

Day	Value
1	151.0383
2	148.2540
3	145.4696
4	142.6853
5	139.9010

2) Analysis report of PM 10 is shown in table 6:-

Table 6: Analysis Report of PM10

Intercept	212.431167
Slope	[-0.10726129]
Mean Absolute Error	55.254617
Mean Squared Error	4725.12157
Root Mean Squared Error	68.739519
R^2, Co-efficient of Determination	0.00059111
R	0.0243129

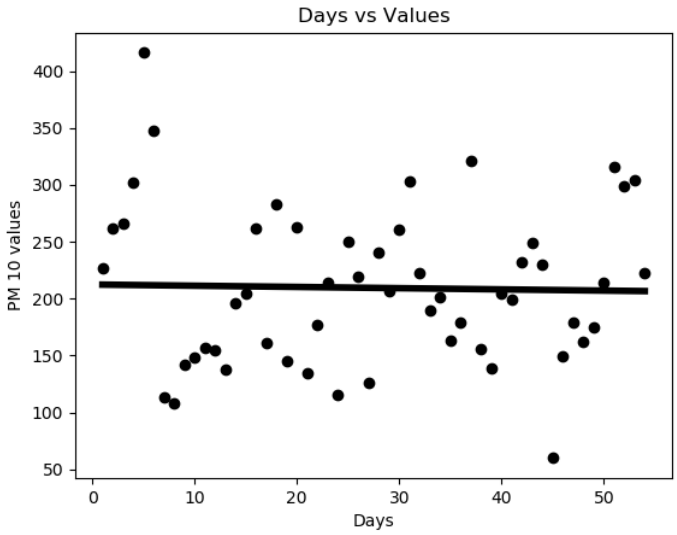


Figure 6 Graph of Days vs PM 10 values

Forecasted values for PM 10 for next five days are shown in table 7:-

Table 7: Forecasted values of PM10

Day	Value
1	206.5137
2	206.4245
3	206.3172
4	206.2100
5	206.1027

3) Analysis report of CO is shown in table 8:-

Table 8: Analysis Report of CO

Intercept	70.810150
Slope	[0.1517371]
Mean Absolute Error	25.225338
Mean Squared Error	1210.77644
Root Mean Squared Error	34.79621
R^2, Co-efficient of Determination	0.005120
R	0.0715586

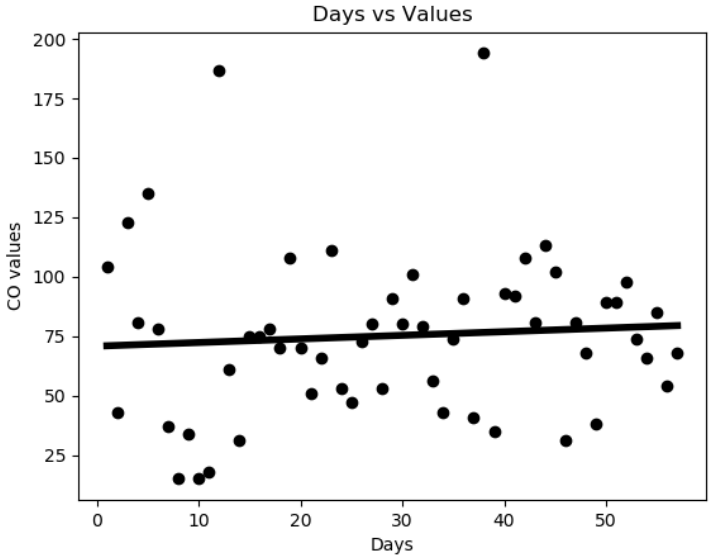


Figure 7 Graph for Days vs CO values

Forecasted values for CO for next five days are shown in table 9:-

Table 9: Forecasted values of CO

Day	Value
1	79.6109
2	79.7626
3	79.9143
4	80.06611
5	80.2178

Similarly, we evaluated the results for remaining air pollutants i.e. NH3, NO2 and SO2 as shown in table 10.

Table 10: Analysis Report of NH3, NO2 and SO2

	NH3	NO2	SO2
Intercept	14.90856	70.52488	37.91833
Slope	[-0.12659]	[0.626056]	[-0.25088]
Mean Absolute Error	4.33206	26.145817	15.09457
Mean Squared Error	37.06594	1249.2006	485.74786
Root Mean Squared Error	6.08818	35.34403	22.03968
R^2, Co-efficient of Determination	0.09187	0.066008	0.03504
R	0.30311	0.256921	0.187195

Forecasted values of NH3, NO2 and SO2 are shown in table 11.

Table 11: Forecasted values of NH3, NO2 and SO2

	NH3	NO2	SO2
Day	**Value**	**Value**	**Value**
1	8.07256	103.70588	23.11615
2	7.94597	104.33193	22.86526
3	7.81938	104.95799	22.61438
4	7.69279	105.58405	22.36349
5	7.56619	106.21010	22.11261

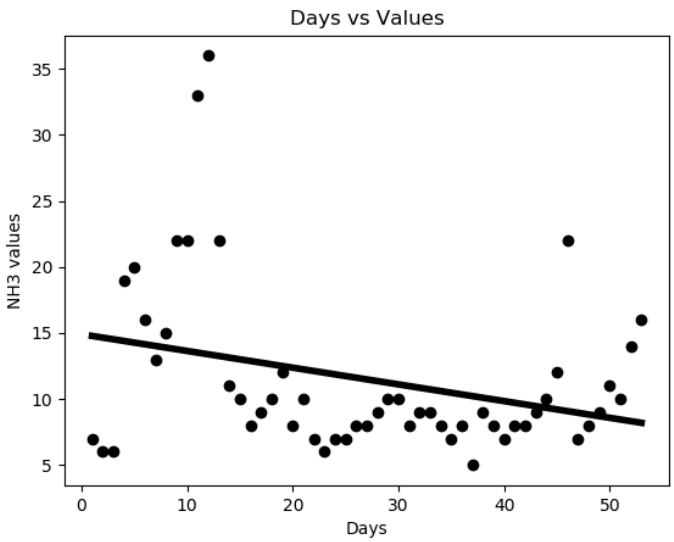


Figure 8 Graph of Days vs NH3 values

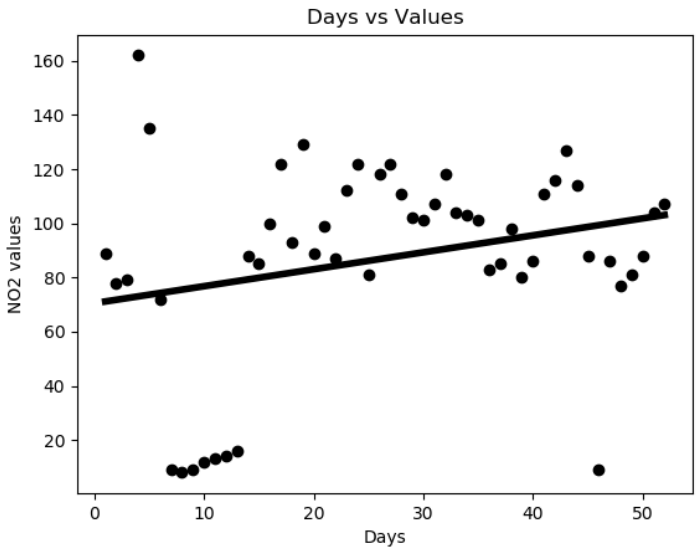


Figure 9: Graph of Days vs NO2 values

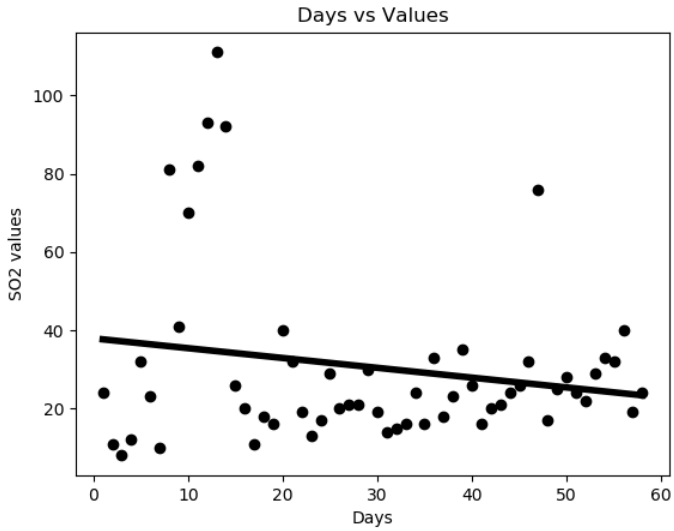


Figure 10 Graph of Days vs SO2 values

It was found that the Ozone O3 pollutant was not found in Anand Vihar area of Delhi but it was found to be the most prominent air pollutant in Kanpur. The concentrations of Ozone were quite high for last 20-30 days in Kanpur due to industrialization. As per our results, the concentrations of air pollutants were quite high for next five consecutive days.

VI. CONCLUSION

The exploration of air quality was deliberated for diverse air contaminants for Anand Vihar area of Delhi. Delhi is one of the most adulterated cosmopolitan cities in India and the foremost malefactors for this are PM 2.5 and PM 10 whereas additional air toxins take account of CO, NO2, NH3 and SO2. The model was assembled using Visual Studio 2017 for linear regression and implemented in Python. The data we used in this was for 50 days and we forecasted the levels of air contaminants for next five days. This will help us to be cautious about the air pollutant concentrations in the atmosphere and to take safety measures against it. The identical prototypical model can be used for soil and water pollution as they are the most conspicuous types of pollution after air pollution. The main goal of this paper was to provide people aforementioned information about air pollutant intensities to be cautious about them.

REFERENCES

[1] Yu, S.; Mathur, R.; Schere, K.; Kang, D.; Pleim, J.; Otte, T. A detailed evaluation of the Eta-CMAQ forecast model performance for O3, its related precursors, and meteorological parameters during the 2004 ICARTT study. J. Geophys. Res. **2007**, 112, 185–194.

[2] Wang, Y.J.; Zhang, K.M. Modeling near-road air quality assessing a computational fluid dynamics model, CFD-VIT-RIT. Environ. Sci. Technol. **2009**, 43, 7778–7783.

[3] Tong, Z.; Zhang, K.M. The near-source impacts of diesel backup generators in urban environments. Atmos. Environ. **2015**, 109, 262-271.

[4] Tong, Z.; Baldauf, R.W.; Isakov, V.; Deshmukh, P.; Zhang, M.K. Roadside vegetation barrier designs to mitigate near-road air pollution impacts. Sci. Total Environ. **2016**, 541, 920–927. Int. J. Environ. Res. Public Health **2017**, 14, 114 18 of 19.

[5] Keddem, S.; Barg, F.K.; Glanz, K.; Jackson, T.; Green, S.; George, M. Mapping the urban asthma experience:Using qualitative GIS to understand contextual factors affecting asthma control. Soc. Sci. Med. **2015**, 140, 9–17.

[6] J. C. M. Pires, S. I. V. Sousa, M. C. Pereira, M. C. M. Alvim-Ferraz, and F. G. Martins, ―Management of air quality monitoring using principal component and cluster analysis—Part I: SO2 and PM10,‖ *Atmospheric Environment*, vol. 42. no. 6, pp. 1249–1260, February 2008.

[7] Ehrendorfer, M. Predicting the uncertainty of numerical weather forecasts: A review. Meteorol. Z. **1997**, 6, 147–183.

[8] Arie Dipareza Syafei, Akimasa Fujiwara, and Junyi Zhang.” Prediction Model of Air Pollutant Levels Using Linear Model with Component Analysis”. International Journal of Environmental Science and Development, Vol. 6, No. 7, July 2015.

[9] Baltazar Frankovic, Viktor Oravec, Ivana Budinska” The Knowledge Modelling of Traffic and Industry Emission from the Air Pollution Control Aspects”. 7th International Symposium of Hungarian Researchers on Computational Intelligence. November 24-25, 2006.

[10] Dr. S. W. A. Ashraf, S. Khanam, A. Ahmad “Effects of indoor air pollution on human health: A micro-level study of Aligarh City-India”. Merit Research Journal of Education and Review Vol. 1(6) pp. 139-146, July, 2013.

[11] Dan Wei” Predicting air pollution level in a specific city”. Stanford University, 2014.

[12] I.N. Athanasiadis, K.D. Karatzas, P. A. Mitkas” Classification techniques for air quality forecasting”. In Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.

[13] Justin R. Chimka , Ege Ozdemir. “A Proportional Odds Model of Particle Pollution”. Environments 2014(mdpi journal), vol. 1, p-54-59, August 2014.

[14] Mihaela M. Oprea.” AIR_POLLUTION_Onto: an Ontology for Air Pollution Analysis and Control”. Artificial Intelligence Applications and Innovations III, Proceedings of the 5TH IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI'2009), Thessaloniki, Greece.P-135-143, April 23-25, 2009.

[15] Ofoegbu E.O.,Fayemiwo M.A, Omisore M.O. “Data Mining Industrial Air Pollution Data for trend analysis and Air Quality Index Assessment using A Novel Back-end AQMS Application Software”. International Journal of Innovation and Scientific Research, Vol. 11 No. 2, pp. 237-247, Nov. 2014.

[16] Alan O. Sykes. "An Introduction to Regression Analysis". Coase-Sandor Institute for Law & Economics Working Paper No. 20, 1993.

[17] Parul Choudhary, Dr. Jyoti Gautam and Nitima Malsa;”Air Quality Prediction using Forward Stepwise Regression by refinement of Ontology with respect to the Indian Domain”, In International Journal of Engineering and Manufacturing Science. ISSN 2249-3115 Vol. 7, No. 1 (2017).

[18] J. Zhang, W. Ding; “Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong”, In International Journal of Environmental Research and Public Health, 24 January 2017.

[19] Arie Dipareza Syafei, Akimasa Fujiwara, and Junyi Zhang; “Prediction Model of Air Pollutant Levels Using Linear Model with Component Analysis”; In International Journal of Environmental Science and Development, Vol. 6, No. 7, July 2015 DOI: 10.7763/IJESD.2015.V6.648.