International Journal of Engineering and Manufacturing Science. ISSN 2249-3115 Vol. 7, No. 1 (2017)

Air Quality Prediction using Forward Stepwise Regression by refinement of Ontology with respect to the Indian Domain

Parul Choudhary [1*], Dr. Jyoti Gautam [2] and Nitima Malsa[3]
[1*] Mtech (CSE), JSSATE, Noida, India
[2] HOD & Associate. Professor. (CSE), JSSATE, Noida, India
[3]Assistant Professor, Department of Computer Science and Engineering, JSSATE, Noida, India

Abstract: Air allows our living planet to inhale – it's the mix of gasses that fills the air, offering life to the planet and animals that make Earth such an exuberant place. However, the gasses you are breathing in could be continuously executing you. According to the World Health Organization, around two million people dies from the impact of contaminated air each and every year. Development of engine auto activity and industry in areas has added to disturbance of the circumstance years ago. Many researchers and foundations address the air contamination control issue. For the assurance of air contamination outflow, dense checking frameworks are working in numerous urban areas. The information from the observing frameworks are utilized via air contamination specialists to take care of the issue and to achieve assignments, for example, here and now forecast of contamination level, settling on choice on crisis counter – measures, and so on. Urban air contamination displaying and information handling systems require elevated affiliation. Utilizing an ontology approach, an intelligible, predictable and non-excess information base might be outlined. In this way, in this work, a regression model for air contamination utilizing ontologies has been executed to anticipate the air quality in a specific city.
Keywords: ontology, air contamination, air pollutant, regression analysis, climate, India

1. Introduction

Air contamination is basically the outside issue perceptible all around—can be fake or happen normally, and are concentrated where people are concentrated. Air contamination has been a peril starting late making honest to goodness risks natural and social success. Government, forces and industry have been at the cutting edge to handle air contamination with the help of technique recreation and mechanical advancement. The fact is to fathom the progression activity in the advancement space and the assorted ways to deal with watch plans in association with scattering of improvement in different inquires about. Common pollution is one of the veritable crisis to which we are facing today. Numerous scientists are tending to the air contamination control issue nowadays. For the assurance of air contamination outflow, thick checking frameworks are working in numerous urban locales. Utilizing manmade brainpower strategies, in particular master frameworks and learning based procedures is extremely encouraging methodology. An information based approach offers another option to the numerical models, coordinating numerous wellsprings of learning in an information base utilized by a deduction motor that can manage vulnerability. Along these lines, in this paper a regression model is proposed utilizing existing air contamination philosophy to predict air contamination quality. The model is made utilizing weka device subsequent to applying information pre-handling procedures.

2. Literature Survey

Arie Dipareza Syafei, Akimasa Fujiwara, and Junyi Zhang[2] In this paper, the expectation of each of air toxins as reliant variable was researched utilizing slack 1(30 minutes previously) estimations of air poisons (nitrogen dioxide, NO2, particulate issue 10um, PM10, and ozone, O3) and meteorological elements and fleeting factors as autonomous factors by considering serial blunder connections in the anticipated fixation. Elective factors choice in light of autonomous segment examination and important segment investigation were utilized to acquire subsets of the indicator factors to be ascribed into the direct model. Utilizing 30-mins interim centralizations of NO2, PM10, and O3, they have shown the impact of different poisons impact and meteorological components. A comparative approach in this paper could be reached out by consolidating days inside week to the information from different stations on different urban areas to build up a forecast.

B. Żogała-Siudem, S. Jaroszewicz. [**3**] The fundamental concentration of research in machine learning and insights is on building further developed and complex models. Notwithstanding, practically speaking it is regularly considerably more vital to utilize the correct factors. One may trust that current prevalence of open information would enable scientists to effectively discover significant factors. However present connected information technique isn't reasonable for this reason since the quantity of coordinating datasets is frequently overpowering. This paper proposes a technique utilizing connection based ordering of connected datasets which can essentially accelerate highlight choice in light of established stepwise relapse strategy. The paper introduces a technique for building relapse show on connected open information at intuitive paces. The technique depends on the utilization of spatial files for effective finding of competitor factors. The strategy has been assessed tentatively on Eurostat information and showed to perform considerably quicker than standard relapse usage.

B. Chandrasekaran, John R. Josephson, V. R.Benjamins[4] The study gives a theoretical prologue to ontologies and their part in data frameworks and AI. The creators additionally examine how ontologies clear up the area's structure of learning and empower information sharing. Hypotheses fall into two general classes: system speculations and substance hypotheses. Ontologies are content

speculations about the sorts of items, properties of articles, and relations between objects that are conceivable in a predefined area of information. They give potential terms to portraying our insight about the space.

Baltazar Frankovic, Viktor Oravec, Ivana Budinska [5] The paper displays an ontological approach for production of air contamination control learning base. The primary reasons why the ontology approach is proposed for demonstrating information base for air contamination control frameworks is that an emblematic learning base can express the space master's learning without the hazard that the understood information will be lost in a tremendous measure of accessible authentic data. The made metaphysics can be reused in numerous comparable application territories and for arrangement of other natural issues.

Dr. S. W. A. Ashraf, S. Khanam , A. Ahmad[**6**] The prior investigation and elucidation in regards to the natural contamination and its effect on the wellbeing, it might be reasoned that the uneven development of populace in city together with congestion is principally in charge of the poor ecological conditions. Subsequent to prior examination with respect to various indoor air contamination and their effect on wellbeing, it might be presumed that the indoor air contamination is influenced by the lodging conditions and living conditions however to some degree it is likewise influenced by the open air condition.

Dan Wei[7] The control of air pollutant levels is quickly getting to be plainly a standout amongst the most essential assignments for the legislatures of creating nations. This undertaking endeavored to apply some machine learning methods to anticipate PM2.5 levels in view of a dataset comprising of every day climate and movement parameters in Beijing, China. The essential objective of the undertaking was the forecast of air contamination level in Beijing City with the round informational index. The best calculation (SVM) gave the 0.722 exactness, 1.000 review and 0.839 F-measure esteem.

I.N. Athanasiadis, K.D. Karatzas, P. A. Mitkas[8] Air quality determining is one of the center components of contemporary Urban Air Quality Management and Information Systems. The paper portrays the correlation work performed between a few measurable strategies and order calculations, on the premise of their execution for particular air quality time arrangement in Athens, Greece. The near investigation of the models execution demonstrated that for the particular experiment the grouping calculations have an extensively better execution contrasted with the factual methods. Future endeavors will focus on joining straightforward arrangement models into accumulated classifiers, (i.e. gathering learning).

Justin R. Chimka , Ege Ozdemir. [10] A linear regression model of molecule contamination and a requested logistic regression model of the important record were chosen for perceptions in the US city of Los Angeles, California. Models were utilized to gauge Air Quality Index (AQI) from a specimen, and were thoroughly analyzed. Linear regression models of AQI through molecule contamination are more supported to foresee direct air quality; requested logistic regression models of AQI straightforwardly are more supported to anticipate great air quality.

Mihaela M. Oprea[11] The paper portrays an ontology for air contamination examination and control, air pollution Onto, and presents its utilization in two contextual investigations, a specialist framework, and a multiagent framework, both committed to checking and control of air contamination in urban regions. The utilization of the ontology in a specialist framework enables the age of the information to base that ought to be reasonable, non-repetitive and finish, while in the event of a MAS it is an imperative help for between operators correspondence.

Mihaela Oprea, M Carbureanu,Elia G Dragomir.[**12**] The paper shows a multi-operator framework, AirQMAS, created for air quality investigation at various stations from the Romanian national air quality observing system. The operators are taking care of the air contamination issues by utilizing cooperative insight and a learning capacity. To think about the communitarian knowledge between various multi-specialist frameworks related to different ecological networks(air, water, soil) keeping in mind the end goal to tackle complex natural issues, for example, the investigation of the effect of a serious air contamination to surface water contamination and soil contamination in basic meteorological conditions, for a particular area.

Natalya F. Noy,Deborah L. McGuinness[14] The control is expand on creators encounter utilizing Protégé-2000 (Protege 2000), (Ontolingua 1997), and (Chimera 2000) as ontology altering situations. In this guide, they utilize Protégé-2000 for their cases. In this guide, the creators have portrayed an ontology improvement strategy for revelatory edge based frameworks. They recorded the means in the ontology improvement process and tended to the mind boggling issues of characterizing class chains of importance and properties of classes and cases.

Niharika,Venkatadri M,Padma S.Rao[15] Air contamination is turning into an ecological danger with the expansion in industrialization and urbanization. This paper concentrates on an exhaustive survey on existing air quality determining systems through delicate registering. In this paper an examination was completed on different Air Pollution/quality forecast methods with most rising delicate processing strategies. The authors watched the fundamental driver for air contamination and the variables that can be capable to limit it.

Ofoegbu E.O.,Fayemiwo M.A, Omisore M.O[16] To build up an air contamination checking application framework for investigating and determining air toxin information with a specific end goal to give data about the nature of air we relax. Air contamination is the presentation of chemicals, particles, organic materials, or other unsafe materials into the world's surface (Wikipedia, 2001). Nigeria is looked with such a significant number of sorts of air contamination, yet the fundamental worry of this examination exertion is air contamination caused by modern discharges. The AQMS application programming plays out the examination of pollutants data to compute air quality file and for estimating information, it enables individuals in a specific area to screen the nature of air they take in. This paper can likewise empower the advancement of other air observing framework in different states/nations to be produced which should be possible through any programming language.

Sameer Kumar, Dhruv Katoria[17] This paper speaks to the distinctive innovations that are utilized in different enterprises and the utilization of various fills that are innate for the arrangement of air contamination. Air contamination has been a hazard as of late posturing genuine dangers to natural and social prosperity. Government, experts and industry have been at the front line to handle air contamination with the assistance of arrangement transformation and mechanical development. The point is to comprehend the advancement action in the innovation space and the distinctive approaches to watch designs in connection to dissemination of development in various locales.

T. Slini, K. Karatzas, A. Papadopoulos[19] This paper depicts the improvement of an application to conjecture the peak ozone levels with the guide of meteorological and air quality factors, in the Greater Athens Area. A linear model has been utilized for the regression and expectation of every day air contamination levels building an operational peak ozone fixation estimating module.

3. Ontology

Ontology significance in logic is "hypothesis of presence or reality"[13]. It expressly indicates the conceptualization of classifications or progressive systems. This is a dialect term used to depict the arrangement of connections made for the semantic web. It is fundamentally a word reference of classes, an asset that figures out what connections can exist and what meaning they have. It is a pattern, a model for developing the structures of data. It advises what has a place with what class. It is a group of information depicting some space for the most part a sound judgment learning area.

Metaphysics can clarify a wide range of surely understood information portrayal structures, some straightforward and others complex. For instance: vocabularies, scientific classifications, and higher request information depiction structures. It tells the use rules. It is rational and reliable shape which conveys the data. It is expressing the diverse purpose of perspectives of clients thus extraordinary ontologies of same thing for various clients like in the event of two book shops which offers books however unique elucidation of same things by two individuals thus two distinct sorts of ontologies are developed for a similar kind of book shops. It is a vocabulary used to depict a specific perspective of some space i.e. the planned significance deciphered.

There are diverse languages utilized for quering an ontology like OWL (Web Ontology Language), OIL(Ontology Interchange Language), and SWRL(Semantic Web Rule Language).

The Ontolgy building devices are Protégé, Apollo, Swoop, Top Braid etc. The different methodologies used for gaining from unstructured data are:

1. Statistical Approach
2. Natural Language Processing Approach
3. Data Mining Approach
4. Web Content Mining Approach
5. Integrated Approach

There are different types of ontologies as explained in the table below

Table 1: Types of Ontologies

Domain Ontologies	Designed to exhibit learning pertinent to an area
Generic Ontologies	Can be applied to variety of domain types(also known as core technology)
Representational Ontologies	Responsible for defining general representational entities without defining what should be presented
Task Ontologies	They provide specific terms for a particular task
Method Ontologies	They provide particular problem solving methods

4. Air Pollution

Air contamination has turned into a danger now-a-days with the expansion in vehicular activity and industry outflow amid recent years. Many creators and organizations address the air contamination control issue. The poisons outflow impacts air quality essentially. The crumbling of air quality makes powerlessness lung contaminations and respiratory illnesses of populace in the locales with high level of air contamination. Auto thickness and industry cause air contamination which relies upon various sorts of autos and industry outflow, on topographical area, meteorological conditions, for example, on temperature, wind, and different variables. For the assurance of air contamination emanation, thick checking frameworks are working in numerous urban locales. The motivation behind the checking framework is to advise specialists about the air quality inside the area. The information from the checking frameworks are utilized via air contamination specialists to take care of the issue and to achieve undertakings, for example, here and now forecast of contamination level, settling on choice on crisis counter-measures, and so forth. The basic leadership depends on quantitative data, got and coordinated from estimations and perceptions, which give data about recurrence and size of progress and whether the built up measures are kept.

Keeping in mind the end goal to give an effective choice help to air quality administration, a scientific model must be made, in view of the connections between ecological assurance and the air poisons (CO, CO2, exhaust cloud, SO2, and so on.). Normally, there are a considerable measure of capricious variables that may impact the level of air contamination, and it is hard to set up with sureness which are the reasons for a lessening or an expansion of an air toxin marker. Customary numerical displaying approaches require overwhelming computational assets and complex information that regularly are not effectively accessible at run time. Utilizing counterfeit consciousness strategies, to be specific master frameworks and learning based procedures, is exceptionally encouraging methodology not just here. These procedures are effectively utilized as a part of different fields of science. The air quality administration frameworks comprise of ecological models and modules for information administration and handling and critical thinking parts. An information based approach offers a contrasting option to the numerical models, coordinating various wellsprings of learning in a learning base utilized by a derivation motor that can manage vulnerability. Utilizing a metaphysics approach, a lucid, reliable and non-repetitive information base might be planned.

5. Regression Analysis

Regression analysis is a factual device for the examination of connections between factors. As a rule, the specialist tries to find out the causal impact of one variable upon another—the impact of a cost increment upon request, for instance, or the eVect of changes in the cash supply upon the expansion rate. To investigate such issues, the agent collects information on the fundamental factors of premium and utilizes relapse to assess the quantitative eVect of the causal factors upon the variable that they impact. The specialist additionally ordinarily evaluates the "factual hugeness" of the assessed connections, that is, the level of certainty that the genuine relationship is near the assessed relationship.

Stepwise linear regression is a strategy for relapsing different factors all the while expelling those that aren't critical.

Forward stepwise regression includes beginning without any factors in the model, testing the option of every factor utilizing a picked show examination paradigm, including the variable (assuming any) that enhances the model the most, and rehashing this procedure until the point that none enhances the model.

5.1 Essential Assumptions & Properties of Regression

As noticed, the utilization of the base SSE measure might be safeguarded on two grounds: its computational accommodation, and its attractive factual properties. We now consider these properties and the suppositions that are important to guarantee them [1]

The theory is that income in "this present reality" are resolved as per the condition I = α + ßE + γX + ε—true values of α, ß, and γ exist, and we want to discover what they are. In light of the noise term e, be that as it may, we can just gauge these parameters.

We can think about the noise term e as an irregular variable, drawn by nature from some likelihood circulation—individuals acquire an instruction and amass work understanding, at that point nature creates an arbitrary number for every person, called e, which increments or reductions salary as needs be. When we think about the noise term as an irregular variable, it turns out to be certain that the appraisals of α, ß, and γ (as recognized from their actual esteems) will likewise be arbitrary factors, in light of the fact that the assessments produced by the SSE model will rely on the specific estimation of e drawn by nature for every person in the informational collection. In like manner, in light of the fact that there exists a likelihood circulation from which every e is drawn, there must likewise exist a likelihood appropriation from which every parameter appraise is drawn, the last dissemination an element of the previous dispersions. The appealing factual properties of relapse all worry the connection between the likelihood circulation of the parameter gauges and the genuine estimations of those parameters.

We start with a few definitions. The base SSE model is named an estimator. Elective criteria for producing parameter gauges, (for example, limiting the whole of mistakes in supreme esteem) are likewise estimators.

Every parameter gauge that an estimator produces, as noted, can be seen as an irregular variable drawn from some likelihood circulation. On the off chance that the mean of that likelihood dissemination is equivalent to the genuine estimation of the parameter that we are endeavoring to appraise, at that point the estimator is impartial. As such, to come back to our representation, envision making a succession of informational collections each containing similar people with similar estimations of training and experience, varying just in that nature draws an alternate e for every person for every datum set. Envision advance that we recompute our parameter gauges for every datum set, along these lines creating a scope of appraisals for every parameter. On the off chance that the estimator is impartial, we would Wnd that by and large we recouped the genuine estimation of every parameter.

An estimator is named reliable on the off chance that it exploits extra information to create more exact evaluations. All the more unequivocally, a steady estimator yields appraises that merge on the genuine estimation of the hidden parameter as the example measure gets bigger and bigger. Subsequently, the likelihood appropriation of the gauge for any parameter has bring down fluctuation as the specimen measure increments, and in the utmost (inWnite test estimate) the gauge will rise to the genuine esteem.

The difference of an estimator for a given example measure is additionally of intrigue. Specifically, let us confine thoughtfulness regarding estimators that are impartial. At that point, bring down fluctuation in the likelihood dissemination of the estimator is obviously alluring — it lessens the likelihood of a gauge that contrasts incredibly from the genuine estimation of the basic parameter. In looking at changed fair estimators, the one with the most reduced fluctuation is named productive or best.

Under specific presumptions, the base SSE standard has the qualities of impartiality, consistency, and productivity—these suspicions and their outcomes take after:

(1) If the noise term for every perception, e, is drawn from a conveyance that has a mean of zero, at that point the entirety of squared blunders standard produces appraises that are fair-minded and predictable. That is, we can envision that for every perception in the specimen, nature draws a noise term from an alternate likelihood dissemination. For whatever length of time that each of these conveyances has a mean of zero (regardless of the possibility that the distributions are not the same), the base SSE foundation is fair-minded and reliable. This supposition is coherently adequate to guarantee that one other condition holds—specifically, that each of the illustrative factors in the model is uncorrelated with the normal estimation of the noise term.

(2) If the circulations from which the commotion terms are drawn for every perception have a similar change, and the noise terms are measurably free of each other (so that if there is a positive noise term for one perception, for instance, there is no motivation to expect a positive or negative noise term for some other perception), at that point the whole of squared mistakes foundation gives us the best or most effective evaluations accessible from any direct estimator (characterized as an estimator that registers the parameter appraises as a straight capacity of the commotion term, which the SSE rule does).

On the off chance that suppositions (2) are damaged, the SSE rule stays unprejudiced and steady however it is conceivable to diminish the difference of the estimator by assessing what we think about the commotion term. For instance, on the off chance that we realize that the fluctuation of the dissemination from which the noise term is drawn is greater for specific perceptions, at that point the measure of the commotion term for those perceptions is probably going to be bigger. Furthermore, in light of the fact that the commotion is bigger, we will need to give those perceptions less weight in our investigation. The factual system for managing this kind of issue is named "generalized least squares".

6. Methodology

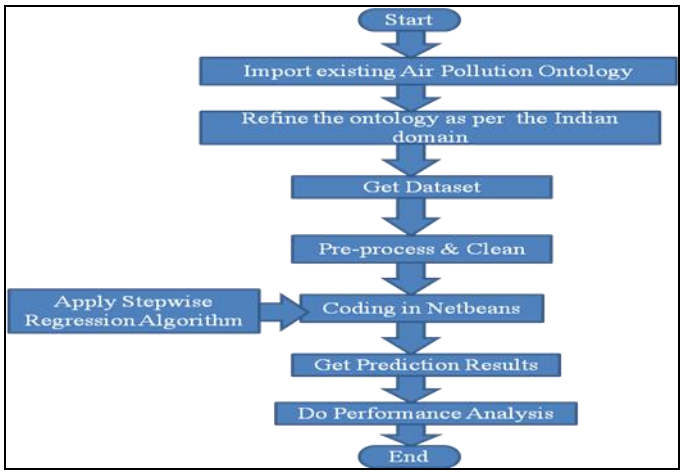


Figure 1: Flow Process Chart

The above chart explains the implementation steps of the project. It is started after importing existing air pollution ontology and refining it according to the dataset. After this the dataset is pre-processed and cleaned for removing noise, outliers etc. And the prediction is calculated after applying Stepwise Regression algorithm and coding in Netbeans. The project is ended after doing performance analysis.

The data for the project is being collected from the following URLs:

- http://newdelhi.usembassy.gov/airqualitydataemb/aqm2013.csv
- http://newdelhi.usembassy.gov/airqualitydataemb/aqm2014.csv
- http://newdelhi.usembassy.gov/airqualitydataemb/jan-nov2015.csv
- http://cpcb.nic.in
- http://cpcbenvis.nic.in/airpollution/database.htm

The coding is done in Netbeans where Weka libraries are imported for regression analysis and Jena libraries are imported for using existing ontology i.e. Air_Pollution_Onto.[11]

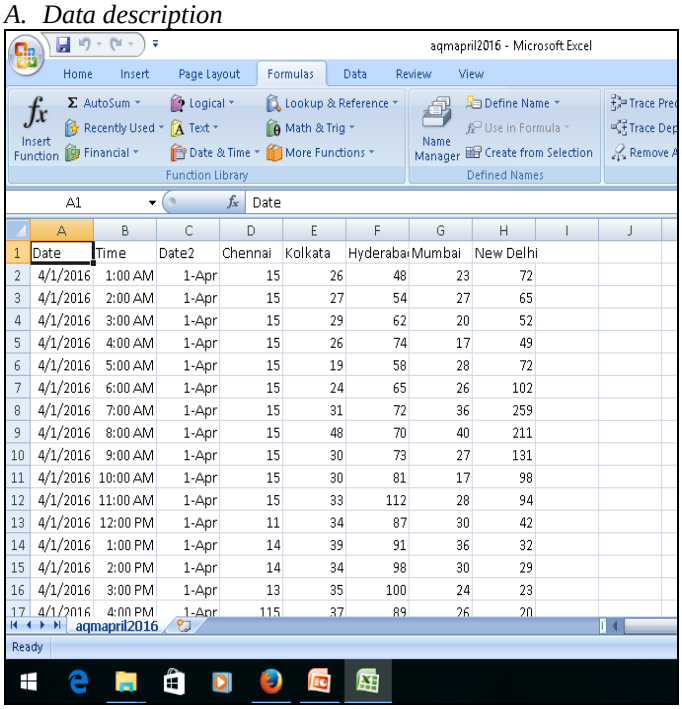
7. Results and Analysis

For daily analysis of air pollution levels, the data for April 2016 shown in Figure 2 is used and a graph is plotted for daily average of five metropolitan cities and another graph is plotted for comparison between Delhi and Mumbai based on daily average of April month.

For monthly analysis of air pollution levels, the Statistics of air pollution is calculated for five metropolitan cities i.e. Delhi, Mumbai, Chennai, Kolkata and Hyderabad and a comparison graph is plotted for these five cities based on the monthly average.

For annual analysis of air pollution levels, the data used are from January 2010 to April 2016. The Statistics is calculated for SO2, NO2 and PM10 air pollution levels for the whole India.

A. Data description



	A	B	C	D	E	F	G	H
1	Date	Time	Date2	Chennai	Kolkata	Hyderaba	Mumbai	New Delhi
2	4/1/2016	1:00 AM	1-Apr	15	26	48	23	72
3	4/1/2016	2:00 AM	1-Apr	15	27	54	27	65
4	4/1/2016	3:00 AM	1-Apr	15	29	62	20	52
5	4/1/2016	4:00 AM	1-Apr	15	26	74	17	49
6	4/1/2016	5:00 AM	1-Apr	15	19	58	28	72
7	4/1/2016	6:00 AM	1-Apr	15	24	65	26	102
8	4/1/2016	7:00 AM	1-Apr	15	31	72	36	259
9	4/1/2016	8:00 AM	1-Apr	15	48	70	40	211
10	4/1/2016	9:00 AM	1-Apr	15	30	73	27	131
11	4/1/2016	10:00 AM	1-Apr	15	30	81	17	98
12	4/1/2016	11:00 AM	1-Apr	15	33	112	28	94
13	4/1/2016	12:00 PM	1-Apr	11	34	87	30	42
14	4/1/2016	1:00 PM	1-Apr	14	39	91	36	32
15	4/1/2016	2:00 PM	1-Apr	14	34	98	30	29
16	4/1/2016	3:00 PM	1-Apr	13	35	100	24	23
17	4/1/2016	4:00 PM	1-Apr	115	37	89	26	20

Figure 2: Showing Excel Sheet of Air Pollutant Level in different cities in ug/m3. No of Records are 720.

A. Daily Analysis of Air Pollution Level

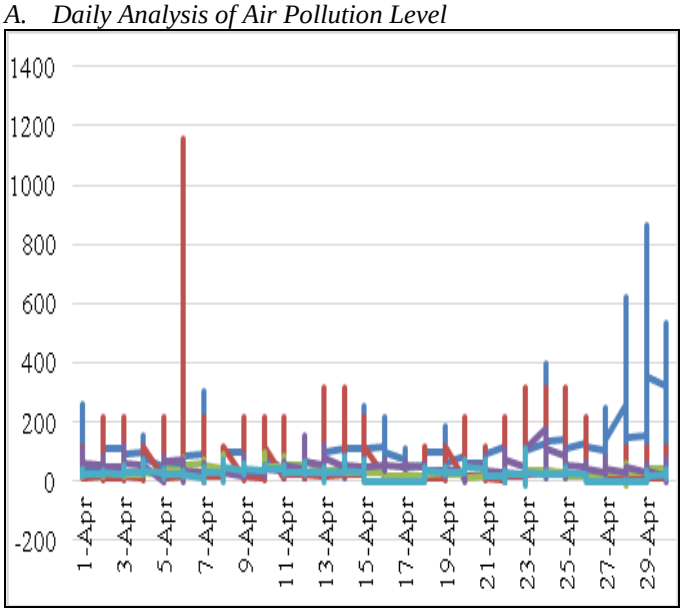


Figure 3: Air Pollution Level on daily average of five metropolitan cities

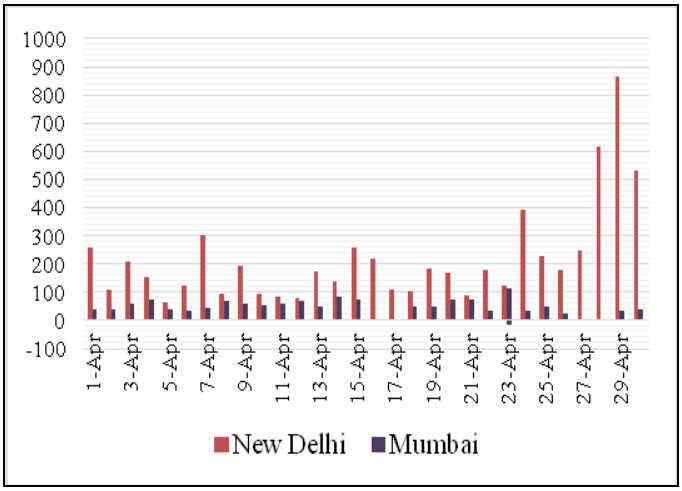


Figure 4: Comparison between Mumbai and Delhi based on Air Pollution Levels on daily average

B. *Monthly Analysis of Air Pollution Level*

Table 2: Air pollution Statistics for five metropolitan cities

Air Pollution Statistics	***Kolkata***	***Hyderabad***	***Delhi***	***Mumbai***	***Chennai***
Mean	34.30417	47.54722222	89.13333333	26.98472222	40.54166667
Standard Error	2.573747	0.797890699	3.352673149	0.689873146	0.598113551
Median	16	47	61	28	37
Mode	15	48	27	0	27
Standard Deviation	69.06088	21.40965 41	89.96166082	18.511239	16.0490707
Sample Variance	4769.405	458.3732885	8093.100417	342.6659693	257.5726704
Kurtosis	101.9533	2.978057183	14.81549358	0.620096831	0.777867407
Skewness	7.811932	0.853557696	3.198714263	0.207548459	0.777426757
Range	1157	177	865	128	117
Minimum	2	0	0	-15	-15
Maximum	1159	177	865	113	102

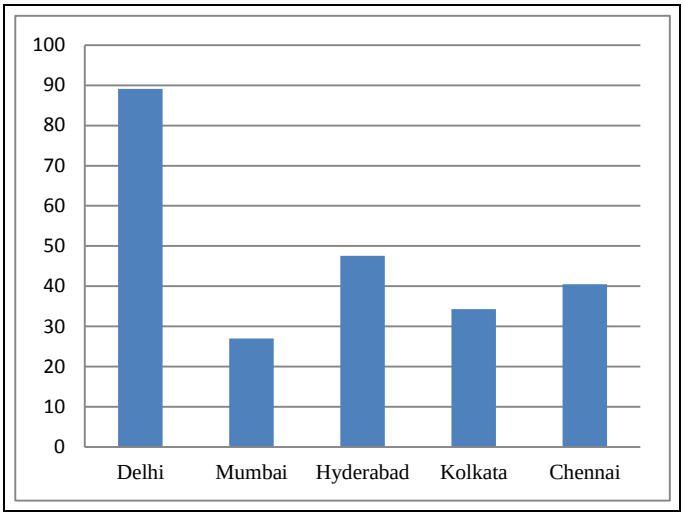


Figure 5: Comparison between 5 cities based on monthly average of Air Pollution Level

According to this graph, Delhi is the most polluted city and Hyderabad is the second most polluted city followed by Chennai in third position and Kolkata in fourth. Mumbai is the least polluted city.

C. *Annual Analysis of Air Pollution Level*

Table 3: Air Pollution Statistics based on annual average

Air Pollutants Statistics	**SO_2**	**NO_2**	**PM10**
Mean	9.696133	23.58011	107.5138
Standard Error	0.619815	1.112131	4.554577
Median	7	20	89
Mode	2	15	76
Standard Deviation	8.338761	14.96219	61.27557
Sample Variance	69.53493	223.8672	3754.696
Kurtosis	7.739552	2.178826	0.974553
Skewness	2.193726	1.422211	1.167368
Range	60	75	308
Minimum	0	0	0
Maximum	60	75	308

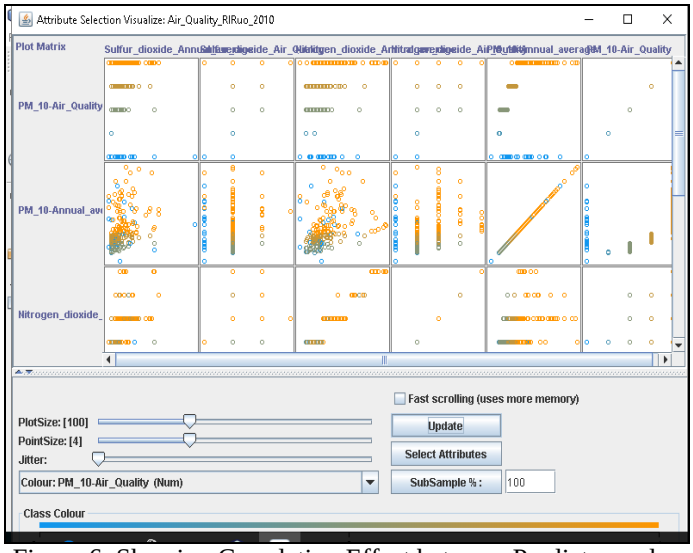


Figure 6: Showing Correlation Effect between Predictor and Target Where on X axis is Predictor and Y axis is Target

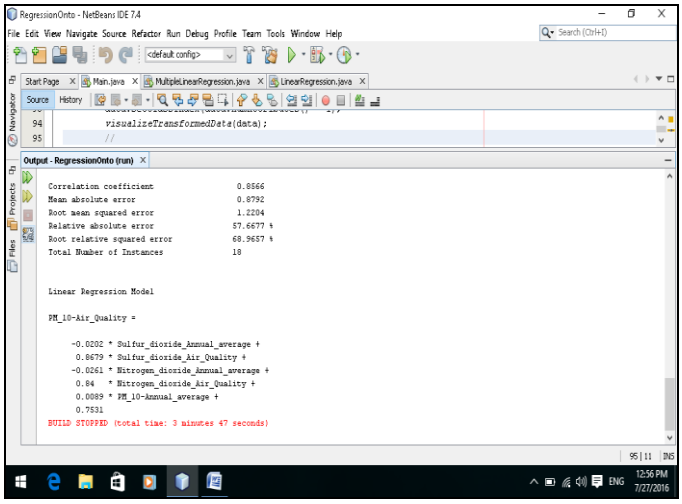


Figure 7: Showing Results of Prediction for PM10

The predictors are SO_2, NO_2 and PM10 air pollution levels. And the targets are SO_2 air quality, NO_2 air quality and PM10 air quality.

Table 4: Total Air Quality Statistics

Air Quality	
Mean	2.906077
Standard Error	0.095863
Median	3
Mode	4
Standard Deviation	1.28971
Sample Variance	1.663352
Kurtosis	0.249551
Skewness	-1.11158
Range	4
Minimum	0
Maximum	4

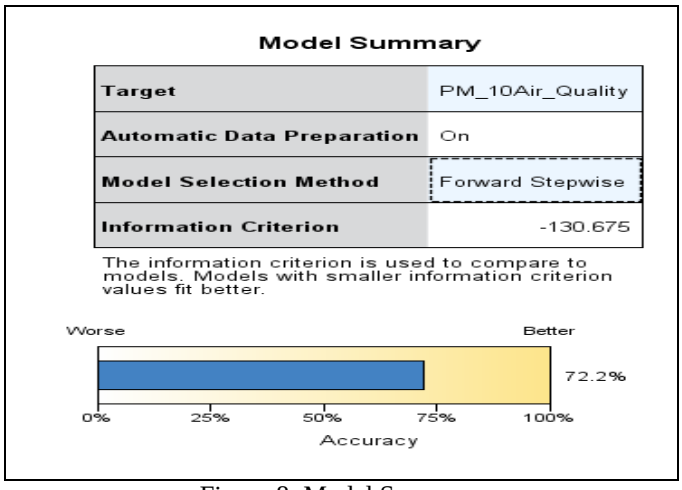


Figure 8: Model Summary

D. Prediction Error for Model Validation

Table 5: Prediction Error and Critical Success Index

Air Quality	Predicted (P)	Observed (O)	Prediction Error(P-O)	Critical Success Index(CSI)
NO2	2.4	2	0.4	0.5
SO2	0.8	1	-0.2	0.5
PM10	3.8	3	0.8	0.5
Mean	2.3	2	0.3	0.5

Lesser the prediction error and more the critical success index means a good model.

CSI = A/A+B+C where A is both predicted and observed value, B is observed value and C is predicted value. Its range is 0 to 1 where 1 means perfect prediction [19].

8. Conclusion and Future Work

The analysis of air quality was discussed for the Indian climate with different air pollutants results of different cities. According to the statistics, Delhi is the most polluted city among five metropolitan cities and PM10 is the biggest culprit of air pollution. A predictor model was made using weka tool for linear regression. The data used is for past 6 years, this can be improved with more data of the past years. The discussed method can be applied for different domain as well like water pollution, noise pollution etc.

References

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

[2] 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

[3] B. Żogała-Siudem, S. Jaroszewicz. “Fast stepwise regression on Linked Data.” In Proceedings of the 1st Workshop on Linked Data for Knowledge Discovery, co-located with ECML/PKDD 2014, pages 17-26, Nancy, France, 2014.

[4] B. Chandrasekaran, John R. Josephson, V. R.Benjamins” What Are Ontologies, and Why Do We Need Them?” IEEE Intelligent Systems and their Applications (Volume: 14, Issue: 1). Page(s): 20 - 26.Jan/Feb 1999.

[5] 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

[6] 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

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

[8] 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.

[9] J. Han, M. Kamber, J. Pei.”Data Mining Concepts and Techniques”. 3rd edition. Published by Elsevier.

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

[11] 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

[12] Mihaela Oprea, M Carbureanu,Elia G Dragomir.” AirQMAS: A Collaborative Multi-agent System for Air Quality Analysis”. ACE 2012

[13] M. Uschold , M. Gruninger” Ontologies : Principles, Methods and Applications”.AIAI-TR-91. Feb.1996.

[14] Natalya F. Noy,Deborah L. McGuinness"Ontology Development 101: A Guide to Creating Your First Ontology". March 2001
[15] Niharika,Venkatadri M,Padma S.Rao" A survey on Air Quality forecasting Techniques". International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 103-107, 2014
[16] 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.
[17] Sameer Kumar, Dhruv Katoria" Air Pollution and its Control Measures". International Journal of Environmental Engineering and Management. Volume 4, pp. 445-450, Nov. 5, 2013.
[18] Stepwise Regression. NCSS Statistical Software. NCSS.com
[19] T. Slini, K. Karatzas, A. Papadopoulos" Regression analysis and urban Air Quality Forecasting: An application for the city of Athens". Global Nest: the International Journal, Vol 4, No 2-3, pp 153 -162, 2002