As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(01): 303-312, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0801052
Available online at: http://www.ijmtst.com/vol8issue01.html

Air Pollution Control using Data Mining

Ashish Singh* | **Manish Ahirwar**

Department of Computer Science and Engineering, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal
*Corresponding Author Mail Id: ashi09singh21@gmail.com

To Cite this Article
Ashish Singh and Manish Ahirwar. Air Pollution Control using Data Mining. *International Journal for Modern Trends in Science and Technology* 2022, *8* pp. 303-312. https://doi.org/10.46501/IJMTST0801052

Article Info
Received: 15 December 2021; Accepted: 12 January 2022; Published: 17 January 2022.

ABSTRACT

The growth and urbanisation[1] of cities has resulted in an expansion in air pollution[2], [3] in recent years. As a result, there has been a lot of analysis and exploration in this field. Pollution is defined as the introduction of harmful materials into the environment. These dangerous components are referred to as pollutants. Contaminants can affect the air, such as lava flows. Human actions, such as rubbish or manufacturing waste[4], can also generate it. Pollutants pollute the air, waterways, and land, wreaking misery on the ecosystem. Every year, an estimated about 7 millions of people die due to the result of environmental pollution all over the world. As according WHO figures, virtually every person on the planet (99 percent)[5] breaths air that exceeds WHO guidelines and includes significant levels of harmful substances, with low- and middle-income countries bearing the brunt of the load. The WHO is aiding nations in their fight against pollutants in the environment.. We utilised data mining to examine current patterns in air pollution in many cities and create future predictions. Regression analysis and multilayer perceptron are indeed the data mining algorithms employed. Different air pollutants such as sulphur dioxide (SO2), nitrogen dioxide (NO2), particulate matter (PM), carbon monoxide (CO), and ozone (O3) have shown patterns. To use the methods described above, we have discovered that the quantity of pollution and contaminants will rise in the next years. In this paper Linear Regression, Decision tree regressor, Random- Forest- Regression and Gradient Booster Regression is used to predict the PM 2.5 concentrations and at what time (month and hour) it is likely to increase. In the data collected for the experiment the hourly data is used which contains the details like concentration of sulphur-dioxide, nitrogen-dioxide, particulate-matter, concentration of rain, atmospheric moisture. Hyper-parameter tuning is also used to enhance the performance of the model..

KEYWORDS: Air Pollution, Pollutants, Data Mining, Machine Learning, Regression.

1. INTRODUCTION

Air quality forecasting is critical for any government's emergency response to excessively contaminated climates. Forecasting not only alert the people to stay away from filthy air, but they also give the administration opportunity to put in place urgent methods to decrease pollution, like reducing the output and emissions of extremely contaminating businesses and prohibiting the use of motorized vehicle[6], [7].It is commonly accepted that metropolitan air pollution does indeed have a significant influence on public health, particularly in poor and developed nations where air quality standards are either unavailable or ineffective. Latest research has already shown significant indications that subjection to air pollution is linked to a variety of illnesses, especially asthmatic and lung congestion. In regards of financial effect, the link among air quality and people's health invariably leads to higher medical expenses, such as hospitalizations and emergency department consultations.The degradation of air pollutants has a substantial influence

on public wellbeing, and it has become a growing source of worry in several nations. Fine particle pollutants, such as PM2.5[8], are particularly simple to take into the lungs, so according to studies, posing a severe hazard to public wellness. Furthermore, according to a recent W.H.O research, air pollution causes around 4.2 millions of mortality per year due to strokes, cardiovascular disease, lung disease, chronic respiratory tracts and other associated disorders.Particulates inside the atmosphere are among the greatest major pollutants with serious health consequences. They include heavy metal ions including mercury, lead, and cadmium, as well as carcinogenic compounds, all of which pose serious health risks. Fuel ashes, petroleum, and diesel engine exhaust pollutants include cancer-causing chemicals such as benzo-pyrene, which can increase the risk of cancer if breathed for a very long time period. The kind of air contamination has changed during the previous forty years. Sulphur dioxides and smoke production associated to earlier smogs have reduced, but the amount of pollution caused by cars has grown. However, air pollution continues to have major medical culmination. At current levels, exposure to man-made particulate air pollution contributes roughly 29,000 deaths per year, estimated to the Govt's Commission on the Medicinal Impacts of Air Pollution. Their current study indicates that Nitrogen oxides pollution produces a similar frequency of premature deaths, but further analysis is required to identify how strongly these impacts are related. The pollution of air is one among a serious danger to physical fitness and also to the habitat, from obscurity that lingered above cities to the burning that took place within the home. Many of people die unexpectedly annually as a consequence of the accumulating effects of environmental (outdoor) and interior air contamination. The bulk of cleaner, low humid air is made up of nitrogen N2 and oxygen O2, which account for seventy eight percent and twenty one percent of the overall amount, accordingly. The remaining one percent is composed up of a mixture of gasses, mostly argon (0.9 percent), as well as trace amounts of atmospheric co2, methane, H2, He, and other components. Liquid water is a common, though very changeable, component of the weather, with concentrations ranging from 0.01 to four percent, in exceptionally humid regions, the moisture percentage of the airflow can exceed upto 5 percent. Some of the pollutants are discussed below :

- **Carbon Monoxide (CO) :**

CO i.e. monoxide of carbon is a very poisonous as well as hazardous contaminant which is notorious for having no distinguishable colour or odour. This gas was once widely used in houses for household heating until it was discovered to be unsafe, and now it has been overtaken by more safer alternatives such as natural gas and electricity. Unfortunately, in human use, this gas is very far from disappearing. Carbon monoxide is most often created by combustion engines that lack sophisticated catalytic converters. Old gas and fuel appliances, incinerators, and even cigarettes are all typical producers of carbon monoxide. Whenever the fuels which are based on the carbon like gasoline, petroleum, gas, oil, timber, and charcoal are burned incompletely (this means that the presence of oxygen is not sufficient), CO is produced. CO, on the other hand, has been mostly produced by vehicle travel in latest years, particularly by gasoline-powered cars. CO may interfere with oxygen transport in the blood and decrease oxygen delivery to the heart, especially in patients with heart disease.

- **SO2 :**

Sulfur dioxide is a colourless gas with a suffocating odour that is produced when sulphur is present as a contaminant in coal or oil. Power stations account for the most bulk of dioxide of sulphur generations, with just a small fraction originating from mobility generators. This noxious gas can cause irritation in the the eyelids and throats, as well as harm lung cells, when inhaled. Dioxide of sulphur is still produced in large quantities by locomotives, ships, aircraft, and other machinery. The gas is also produced by industrial operations, notably mineral mining and purifying. One sulphur atom and two oxygen atoms make up SO2. When it interacts with other chemicals, it produces hazardous molecules like sulphuric acid (H2SO4), which may create acid rain. SO2 causes asthmatics' airways to constrict as a result of the airway's immunological reaction to the irritant. Sulfur Dioxide, like Nitrogen Dioxide, is a molecule that frequently reacts with other pollutants in the air to produce toxic

acids, but it is also dangerous on its own. Respiratory difficulties, vision problems, and even heart and circulation problems are all common side effects of Sulfur Dioxide poisoning. Sulfuric acid, which is produced by combining sulphur with water, is a powerful acid that is harmful not only to humans but also to plants, soils, and water systems.

- **Particulate Material (PM) :**

Particulate Matter[9] is made up of a variety of particles, some of which exist naturally and others which are manufactured. Sand and sea salt are occurs naturally particles, whereas others are produced by chemical processes in the atmosphere or released by traffic pollution. Soot, grime, and chemical by-products generated by combustion or chemical mixing make up the majority of the material. Particle matter is created by almost every chemical and fuel-related operation, as well as more mundane ones like farming and road building. Particulate matter is produced by any process that creates a physical by-product. Particle Matter may be both irritating and hazardous, obstructing vision on the road and creating respiratory issues. Particulate Matter has been linked to a variety of heart, lung, and eye diseases in people, as well as an increased risk of getting cancer later in life. Long-term particle exposure (specifically PM2.5) has been linked to an increased risk of death, particularly from heart and lung diseases. According to recent research, excessive levels of PM2.5 in children can permanently damage lung function. Asthma patients may be affected by high particle levels. It is classified based to its micrometre size. **PM10** denotes particles smaller than 10 micrometres, sometimes known as the 'coarse fraction. **PM2.5** refers to particles with a diameter of less than 2.5 micrometres, also known as the "fine fraction. PM2.5 is believed to be more harmful to human health. All types of vehicles, as well as various industrial operations and forest/wildfire burns, produce these tiny particles. PM10 particles comprise PM2.5 particles as well as coarser particles such as dust, pollen, and mould. These particles can come from a variety of sources, including road travel, diesel trains, cargo, industry, and solid fuel combustion, as well as naturally occurring materials like sand and sea salt.

- **Nitrogen Oxide(NOx):**

Nitrogen oxides, or NOx, are a group of extremely reactive, toxic gases produced whenever fuel is burnt at extreme temps. Automobiles, as well as industrial emissionsincluding power stations, commercial boiler, cement factories, and turbines, generate NOx emissions, which are reddish brown. In the atmosphere, nitrogen oxides have troublesome chemical interactions with volatile organic molecules. On hot summer days, these processes generate haze. NO and NO2 are the two most common nitrogen oxides**.** It's a secondary pollutant that forms when nitric oxide (NO), which is produced during the burning mechanism, interacts with oxygen in the air. NO2 is a suffocating and obstructive gas in the air that reacts with the other factors to generate nitric acid and organic nitrates, leading to acid rain production. Nitrogen Oxide, as you might expect, has a significant impact on humans, increasing the risk of respiratory disorders, cancer, and other lung issues. Acid rain, which is caused by the release of nitrogen dioxide into the atmosphere, is very damaging to plants and animals all over the world.

- **O3:**

Ozone is a colourless, light blue gas consisting of three oxygen atoms bonded collectively. At high quantities, it produces a unique odour. Ozone is a crucial constituent of photo - chemical haze, and it is produced by a complex interactions between dioxide of nitrogen and hydrocarbon under the effect of sunlight. Ozone is basically found naturally gas that shelters the planet from the sun's damaging UV radiation approximately ten to thirty miles well above earth 's crust in the upper levels of the environment. Ground-level ozone is produced by biochemical mechanisms among a variety of airborne oxides, that, whenever subjected to light from the sun, can produce new compounds and release Ozone as a by-product. The obvious culprits, such as automobile exhaust, factory processes, electric utilities and power plants, and even some chemical solvents, account for the bulk of Ozone-forming emissions. Ozone is produced by chemical interactions in light from the sun among pollutants from manufacturing sites, car emissions, and organic chemical solvent products. When oxides of

nitrogen and volatilis organic components react in the sunlight, ozone is formed. Because most living species are poisoned by ozone, a rise in ground-level ozone has a significant impact on human health as well as the general health of numerous ecosystems on land and at sea. Ground-level ozone is hazardous to one's health, activating asthma, creating breathing issues, lowering lung function, and possibly developing lung illnesses. Ozone is also damaging to plants and trees at high enough concentrations. Ozone may degrade construction components, sculptures and memorials, and natural stone features in the environment.

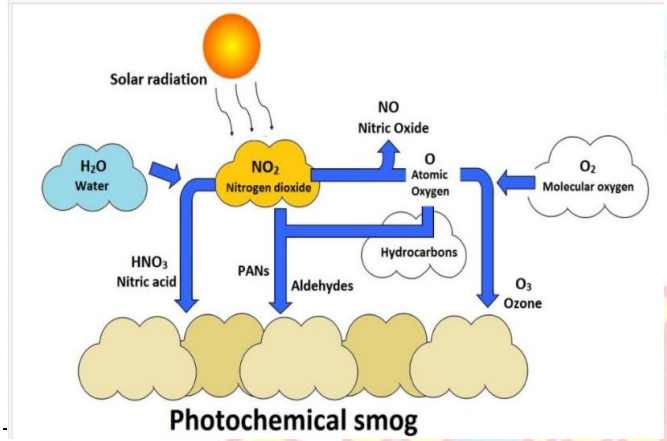

Figure 1:- Pollutant Cycle.

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health warnings of emergency conditions. The entire population is more likely to be affected.
Hazardous	301 to 500	Health alert: everyone may experience more serious health effects

Figure 2:- Pollutant scale.

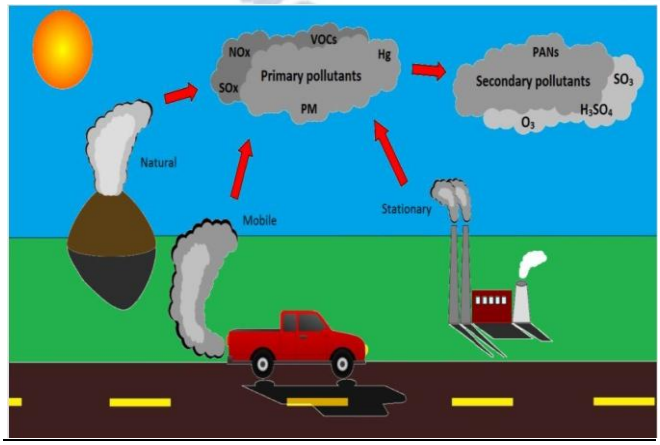

Figure 3:- Formation of pollutants.

II. RELATED WORK

[10]The suggested approach will undoubtedly assist in enhancing air pollution forecasting in our growing cities. Forecasting of Multivariate Multistep Time Series The use of the Randomized Forest approach improves the effectiveness of the air pollution forecasting model while reducing its intricacy. We're also employing a feature selection strategy to improve our forecast.

[6]This research uses the IVE methodology to offer a novel grid-based movable source emissions inventories. For the years 2008–2009, CO, NOx, and PM automobile exhaust in Delhi were around 509, 194, and 15 tonnes per day, respectively. The greatest CO, NOx, and PM emissions are released during the start-up phase, accounting for roughly 86 percent, 27 percent, and 71 percent of total emissions, respectively. CO and PM emissions are mostly produced by gasoline and diesel automobiles. CNG cars are a significant source of NOX emissions.

[11]The goal of this research was to employ time series regression forecasting and data mining approaches to decipher the varied trends in distinct kinds of contaminants. The R studio platforms and visuals were used to forecast future concentrations of pollution in Delhi using the R programming language. As a consequence, the rising trend in NOX in the years ahead may be linked to an increase in the amount of automobiles on the streets, emissions from neighbouring industrial zones, and the activities of thermal power plants. PM10 and PM 2.5 levels have risen due to kerb-side dust, building operations, and agricultural stubble burning in neighbouring states. Smog pollution is also caused by increased NOX, particle matter (PM10 and PM2.5), and ground level O3.

[12]Using time series datamining methods, this research provides an effective approach for predicting the concentrations of various air contaminants. For the prediction of air contaminants, the time series datamining technique CTSPD was utilised. When the suggested solution's forecasts are evaluated to SAFAR-predictions, India's it is discovered that the suggested solution produces more accurate results. It was discovered by analysing the observed air quality patterns that now the level of a contaminant does not

have to be dependent on the concentrations of all of the associated pollutants. By employing air quality data from several places as the data source, this problem may be expanded to incorporate the spatial properties of the data.

[1]Utilizing an innovative device platform, datas collected are evaluated and then utilized to predict contaminant concentration levels. The platform learns from the collected data to develop prediction model using machine learning-based algorithms. This models forecast concentration levels one, eight, twelve, and twenty-four hours ahead of time. Due to the obvious tree structure effectiveness and excellent generalisation ability, M5P surpasses other algorithms for all gases in all horizons in terms of NRMSE and PTA, according to numerous studies. In our situation, however, ANN performed poorly because to its weak generalisation capacity when working with tiny datasets with many features, resulting in a complicated network that overfits the data, whereas SVM performed better than ANN due to its flexibility with high-dimensional data.

[9]A forecasting framework including meteorological parameters based on MKL is provided in this research for PM2.5 prediction in the near future. The MKL approach is being used for the first time in PM2.5 forecasting. Furthermore, we present a first-order primal-dual alogrithm for solving the MKL issue that does not require a line search of step size. The suggested prediction paradigm incorporates both linear and nonlinear relationships among PM2.5 mass concentrations and meteorological factors,, according to experimental results.. On the other hand, we discover that automated factor choosing via weighted kernels is contained in MKL, which is not accessible in traditional SVR, based on the computation outcomes of kernel weights.

[2]CO is mostly released by the transportation and home sectors, according to research. The transportation industry contributes the most NOx, followed by power plants and the home sector, with an upward trend from 2000 to 2010. Power plants and the transportation sector contribute the most to PM10 emissions. PM10 emissions from the transportation sector have decreased dramatically.

[8]The framework provided here is a new method for identifying and classifying locations based on their PM2.5 composition. The 31 clusters found were divided into four categories, with 21 clusters having two or more sites. The proposed method is reasonably resistant to reliability based at collecting locations as well as the selection of locations to consider. The eastern United States are home to the clusters during the first category. For N, V, Si, Ca, Mn, and Cr, they have lower than average saturation factors. The metropolitan and remote locations, on the other hand, are forming different clusters. SO4 2, Se, and As concentrations range from medium to high at these locations. EC enrichment varies from average to low depending on urbanisation, while OC enrichment is average.

[13]For building the air quality– prediction algorithms at the ancient landmark, Taj Mahal, Agra, regression along with neural network having multi–layer perceptron were applied. It provides superior air pollution prediction strategies based on ANN methodologies for modelling periodically hourly time series data with PCA methods performing superior than MLR methods at Taj Mahal, Agra. In order to pick the important contributing variables, a novel technique focused on generalized linear models, PCA, was used. The model is based on climatic factors and amounts of air contaminant concentrations that have been observed. The anticipated and actual pollution amounts at Taj Mahal, Agra, were found to be in fair compliance. Because of the inputs and design of the system itself, both suggested models were shown to be unable to predict adequately during highly concentrated pollutant periods.

[14]This study describes an air quality surveillance system that uses several gas sensors and uses a worldwide positioning systems module to identify the contaminated region. The method collects air pollution readings in big cities using public transportation buses. This approach will instantly assist a large number of individuals. Asthmatic patients, joggers, and others worried about air quality may be among them. The enormous volume of studied data may aid government agencies that monitor and apply pollution regulations, resulting in a greater knowledge of the various contaminants that harm the metropolitan ecosystem.

It's also useful for air pollution control and locating big contaminating factors in different parts of a city.

Table 1. Comparison of methods implemented to monitor Air pollution

Features	**K.ThanweerBasha**[15]	**Poonam Pal**[16]	**Dr. M. Newlin Rajkumar**[17]	**Lalit Mohan Joshi**[18]	**Rose Sweetlin .T** [19]	**RiteekaNayak**[20]
Application Area	industries sector	To monitor the air quality of environment	Efficient log management	Roadside, industrial pollution Monitoring	vehicle emission problem	Air quality monitor system
Algorithm	NA	information fusion algorithm	SMS based algorithm	NA	Spectrophotometry method	NA
Technique Used	IOT	IOT	GPS	RFID	Spectrophotometry,	
Microcontroller	Arduino UNO board	Arduino UNO	Raspberry pi	AVR UNO, XMEGA 2560 UNO	Raspberry Pi 3	Arduino Uno R3 microcontroller
Type of Sensor	Sensors CO2, O2, Methane, H2, Ammonia Hydrogen Sulfide –	MQ135 gas sensor	CO2 sensor	DHT11 , dust, gas sensor	Linear image sensor	*MQ135 Sensor*
Technology	Cloud	Bluetooth	GPS	VoIP applications and Bluetooth	My SQL	ESP8266 Wifi
Hardware	Arduino UNO board, Breezo Meter, wifi module, LCD	Arduino UNO board, buzzer, wi-fi module ESP8266	WSM, CO2 sensor, LCD	AVR UNO, ESP8266,buzzer, wi-fi module,LCD	Raspberry Pi 3, buzzer, LCD	Arduino Uno R3, LCD, ESP8266, MQ135
Software	MATLAB	NA	MAQU MON	firmware	MySQL, CSS SOFTWARE	COM3
Air Pollution Parameters	Dust particles	Poison gases	CO2 gas	Sound & dust particles	Vehicle emission	CO2 gas, dust particles
Air Pollution Behaviour Information	Yes	NO	Yes	Yes	yes	Yes
Reliability	Higher	lower	Moderate	Lower	Lower	Lower
Speed	Higher	Higher	Moderate	Higher	Lower	Moderate
Advantaged	Awareness towards the air pollution	Better quality of PPM on LED screen	Improving the quality	Industries parameter monitoring	Data easily collected	Easy to monitor
Disadvantages	complicated	Low costing	Complexity	costly	Time taken	Time taken

Problem Statement

The In the domain of air pollution the main issue is that the data coming from all the sources are not accurate there are some missing data, some null values, some garbage data etc. The works which has been performed previously has given less accurate results. The prediction which is being done gives us less accuracy.

III. METHODOLOGY

In this paper we have used machine learning algorithm and data mining techniques to predict the concentration of the PM2.5 in the air. We have also compared the outcomes of different regression techniques. The purpose of this study is to use publicly accessible Bhopal weather data to determine whether we can estimate the quantity of PM2.5 in the air based on other environmental factors using machine learning techniques. I'm hoping to develop a prediction model that is both accurate and has a low Root Mean Square Error (RMSE).

Process Description

The following diagram makes it easier to understand how we proceed.

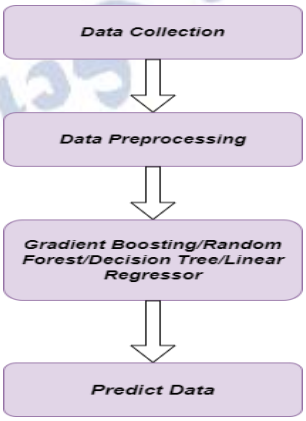

Figure 4:- The flow chart showing the overall steps.

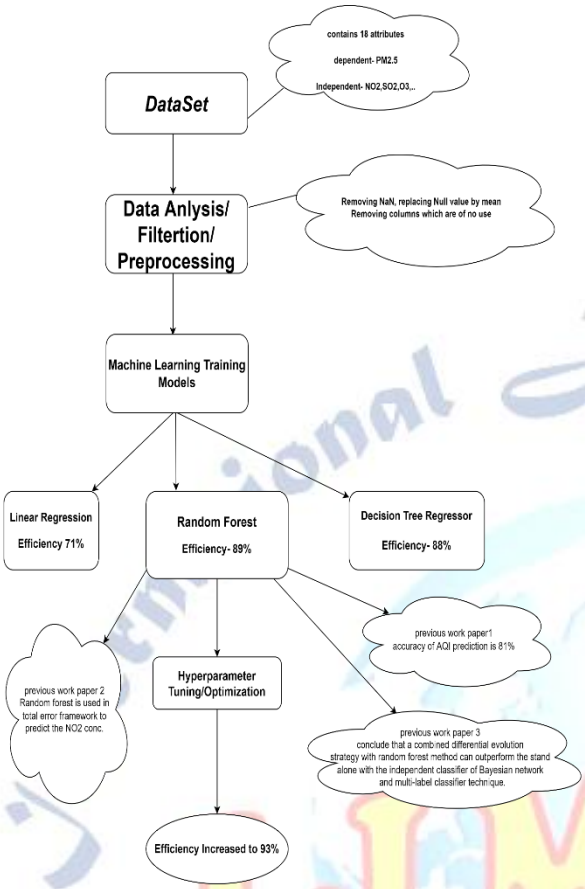

Figure 5: Flow Chart Of the Work

Dataset Information :

For the prediction to be done the datas of every hour from several nationally managed locations for monitoring the quality of air are utilised in this collection of data. The Pollution control board Bhopal provided the air-quality data. Meterological datas coming from every monitoring station is compared with the weather office nearby. NA stands for missing data.

Data Mining :

Data mining[4],[21], [22],[23], [24] is a technique for transforming unstructured converting data into useful knowledge. Organizations may gain a better understanding of their customers by using software to look for trends in large amounts of data. This enables them to create more effective advertising strategies, increase sales, and save costs. Efficient data gathering, warehousing, and computer interpretation are all required for data mining. Data mining is the process of autonomously examining enormous amounts of data for patterns and trends which go further than basic comparison. Data mining estimates the likelihood of upcoming occurrences by utilising advanced mathematical algorithms for data segments. Data mining is also known as data knowledge discovery (KDD)[5].Data Mining is related to Data Science, which is done by a professional in a given circumstance, on a given data collection, and with a certain goal in mind. Text mining, online mining, audio and video mining, graphical data mining, and social media mining are only some of the services available. It's done using either basic or extremely specialized software. By outsourcing data mining, all of the work may be completed more quickly and at a lower cost. Specialized businesses can also take use of new technology to acquire data that would otherwise be hard to locate manually. Although there is a wealth of material available on multiple platforms, there is a scarcity of expertise. The most difficult task is to evaluate the data in order to extract significant information that may be utilised to solve an issue or advance the firm. There are a plethora of strong tools and approaches for mining data and extracting more information from it.

Machine Learning :

Machine learning is a form of artificial intelligence (AI) that enables systems to understand and grow under their self without the need for programming. The construction of computer programmes that can collect data and adjust on their own is what machine learning is all about. The training process begins with observations or data, such as examples, firsthand experiences, or instructions, so that we may look for patterns in the data and make better decisions in the future based on the examples we provide. The overarching objective is for computers to learn on their own, without the need for human intervention, and to adapt their behaviour as a result.

Linear Regressor :

Linear regression is the most basic and extensively used type of predictive analysis. Regression's purpose is to look at two things: (i.) Can a collection of predictor variables be used to anticipate an outcome (dependent) variable? (ii.) Which factors in particular are significantly predictive of the outcomes measure, and how do these impact it (as seen by the magnitude and sign of the β forecasts)? The purpose of these regression

estimates is to show how one dependent variable interacts with one or more independent variables. The simplest version of the regression equation with one dependent and one independent variable is y = c + b*x, where y represents the estimated dependent variable score, c represents the constant, b represents the regression coefficient, and x represents the independent variable score.

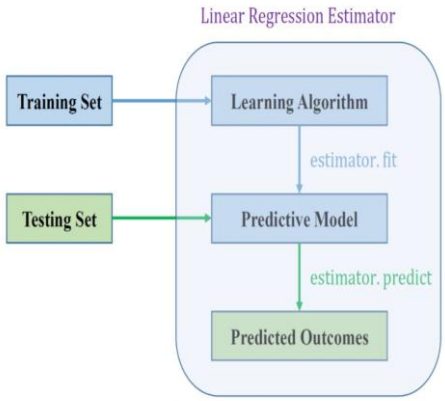

Figure 6:- Figure shows how the training and testing datasets are given to the model for learning and prediction.

Random forest Regressor

This is a prominent machine learning technique that is classified as ensemble learning, implying that it is a mixture of several Decision Tree classifiers. Since it is an ensemble of classifiers, its primary goal is to improve the model's performance in terms of accuracy. As is commonly said, increasing the decision tree classifier leads to increased predictive performance and eliminates the concern of overfitting. A Random Forest is an aggregation methodology that solves regression and classification issues using many decision tree structures using a process called Bootstrap and Aggregation, often referred as bagging. Rather of relying on individual decision trees, the basic concept is to combine several decision trees to arrive at a final result. Random Forest is a foundational learning paradigm that employs many decision trees. The dataset is sampled at random for row and feature sampling, resulting in sample datasets for each model. The bootstrap is the name for this part. To produce a more precise forecasting than a solo model, the ensemble learning technique combines predictions from numerous machine learning algorithms.

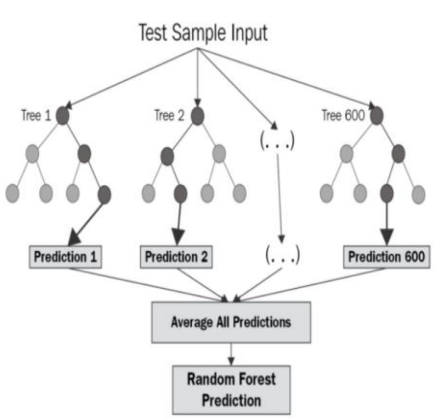

Figure 7:- Above Diagram shows the working of a Random Forest Classifier.

Decision Tree Regressor

This is a guided Machine Learning method that may be used to do both classification and regression. It's a tree-structured classification in which the interior nodes reflect dataset characteristics, the tree's branching provide judgment rules, and the leaflets mostly represent the conclusion of decisions made based on the rules provided. In the shape of a tree structure, Regression or classification models are built using a decision tree. It continually breaks apart a dataset over smaller and smaller portions also while building a decision tree. The ultimate output is a tree with decision nodes and leave nodes. A decision node (such as Outlook) can contain two or even more branching, each representing a value for the property being checked. A leaf node is a representation of a numerical goal option. The best predictor is represented by the root node, which is the uppermost decision node in a tree. Decision trees can handle both categorical and arithmetic data. In decision processes, a decision tree could be utilized to graphically & clearly depict decisions and decision-making. In data mining, a decision tree is used to explain data.

Gradient Boosting Regressor

A prominent boosting technique is gradient boosting. Every prediction in gradient boosting rectifies the mistake of its previous. Unlike Adaboost, the learning instance values are not adjusted; rather, each prediction is trained using the predecessor's leftover errors as label. CART is the foundation learner in a method called Gradient Boosted Trees. Gradient boosting is a machine learning approach that may be used for a

variety of applications, including regression and classification. It returns a prediction model in the form of an ensemble of weak prediction models, most often decision trees.The resultant approach is called gradient-boosted trees when a decision tree is the weak learner; it generally beats random forest. A gradient-boosted trees model is constructed in the same stage-wise manner as other boosting approaches, but it differs in that it allows optimization of any differentiable loss function.

Hyperparameter Optimization

The task of selecting a collection of ideal hyperparameters for a learning algorithm is known as hyperparameter optimization or tuning in machine learning. A hyperparameter is a value for a factor that is used to influence the process of training. Other factors, such as nodes weights, are, on either hand, learnt. To generalise diverse data trends, the very same machine learning model may demand different constraints, weights, or learning rates. These values are known as hyperparameters, and they must be fine-tuned in order for the model to perform the machine learning task optimally. Hyperparameter optimization identifies a tuple of hyperparameters that results in an optimum model that minimizes a predetermined loss function on independent data.

Result:

In the research work we have compared the efficiency of prediction of PM2.5 by the random forest regressor, Decision tree regressor, Linear regressor and gradient boosting regressor and after comparing all these we got to the result that Random Forest regressor has the best efficiency and further on modifying the algorithm we got the output with efficiency 93.43% accuracy

Table 2: Results Obtained

Model	Accuracy
Linear Regressor	71.38%
Decision Tree Regressor	87.97
Random Forest Regressor	89.04%
Random Forest Regressor with Hyperparameter Optimization	93.43
Gradient Boosting Algorithm	87.65

IV. FUTURE SCOPE AND CONCLUSION

In the research work we have concluded that the pollution specially air pollution has became one of the major threat to humans. We need to control the air pollution using these predictions to raise a risk alert so that the after exceeding the level certain measures must be taken by the government to stop the pollution. One thing we should notice that it is not the government who can stop the risk of pollution but we the peoples should be aware and we should take certain measures or steps which would help in reducing the pollutants so that we can live a healthy life Furthermore coming on this research part the efficiency which has been achieved in can be improved in future if the dataset are consistent and appropriate and if some more refinement can be made in the algorithm part.

REFERENCES

[1] K. B. Shaban, A. Kadri, and E. Rezk, "Urban air pollution monitoring system with forecasting models," IEEE Sens. J., vol. 16, no. 8, pp. 2598–2606, 2016, doi: 10.1109/JSEN.2016.2514378.

[2] H. Khas, "Assessment of gaseous and respirable suspended particulate matter (PM 10) emission estimates over megacity Delhi : Past trends and Future Scenario (2000-2020) Rati Sindhwani *, P . Goyal , Saurabh Kumar Anikender Kumar," pp. 1–6, 2014.

[3] S. Taneja, N. Sharma, K. Oberoi, and Y. Navoria, "Predicting trends in air pollution in Delhi using data mining," India Int. Conf. Inf. Process. IICIP 2016 - Proc., pp. 1–6, 2017, doi: 10.1109/IICIP.2016.7975379.

[4] Y. Ma, M. Richards, M. Ghanem, Y. Guo, and J. Hassard, "Air pollution monitoring and mining based on sensor Grid in London," Sensors, vol. 8, no. 6, pp. 3601–3623, 2008, doi: 10.3390/s8063601.

[5] F. T. Souza and W. S. Rabelo, "A data mining approach to study the air pollution induced by urban phenomena and the association with respiratory diseases," Proc. - Int. Conf. Nat. Comput., vol. 2016-Janua, pp. 1045–1050, 2016, doi: 10.1109/ICNC.2015.7378136.

[6] P. Goyal, D. Mishra, and A. Kumar, "Vehicular emission inventory of criteria pollutants in Delhi," Springerplus, vol. 2, no. 1, pp. 1–11, 2013, doi: 10.1186/2193-1801-2-216.

[7] A. S. Nagpure, K. Sharma, and B. R. Gurjar, "Traffic induced emission estimates and trends (2000-2005) in megacity Delhi," Urban Clim., vol. 4, pp. 61–73, 2013, doi: 10.1016/j.uclim.2013.04.005.

[8] E. Austin, B. A. Coull, A. Zanobetti, and P. Koutrakis, "A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition," Environ. Int., vol. 59, pp. 244–254, 2013, doi: 10.1016/j.envint.2013.06.003.

[9] X. Xu, "Forecasting air pollution PM2.5 in Beijing using weather data and multiple kernel learning," J. Forecast., vol. 39, no. 2, pp. 117–125, 2020, doi: 10.1002/for.2599.

[10] H. Patel and S. Saket, "Air Pollution Prediction System for Smart City using Data Mining Technique: A Survey," Int. Res. J. Eng. Technol., pp. 990–995, 2019, [Online]. Available: www.irjet.net.
[11] N. Sharma, S. Taneja, V. Sagar, and A. Bhatt, "Forecasting air pollution load in Delhi using data analysis tools," Procedia Comput. Sci., vol. 132, pp. 1077–1085, 2018, doi: 10.1016/j.procs.2018.05.023.
[12] M. Yadav, S. Jain, and K. R. Seeja, Prediction of air quality using time series data mining, vol. 56. Springer Singapore, 2019.
[13] D. Mishra and P. Goyal, "Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra," Atmos. Pollut. Res., vol. 6, no. 1, pp. 99–106, 2015, doi: 10.5094/APR.2015.012.
[14] U. Lanjewar and J. Shah, "Air pollution monitoring & tracking system using mobile sensors and analysis of data using data mining," Int. J. Adv. Comput. Res., vol. 2, no. 4, 6, pp. 19–23, 2012.
[15] "Awareness of Air Pollution through IoT | Connected Papers." https://www.connectedpapers.com/main/2b3f512e2cad76f2fb18fb4c89d3d299acef3574/Awareness-of-Air-Pollution-through-IoT/graph (accessed Dec. 30, 2021).
[16] P. Pal, R. Gupta, S. Tiwari, and A. Sharma, "IOT BASED AIR POLLUTION MONITORING SYSTEM USING ARDUINO," pp. 1137–1140, 2017.
[17] M. N. Rajkumar, "IOT Based Smart System for Controlling Co 2 Emission," no. June 2019, 2017, doi: 10.13140/RG.2.2.26703.33444.
[18] L. M. Joshi, "Research paper on IOT based Air and Sound Pollution Monitoring System," vol. 178, no. 7, pp. 36–49, 2017.
[19] R. S. T, "Control of vehicle pollution through Internet of things (IOT)," vol. 3, pp. 144–147, 2017.
[20] "IoT Based Air Pollution Monitoring System | Connected Papers." https://www.connectedpapers.com/main/fa1c11e5f710dabff12a20f3dbc9f2473fa45782/IoT-Based-Air-Pollution-Monitoring-System/graph (accessed Dec. 30, 2021).
[21] K. Siwek and S. Osowski, "Data mining methods for prediction of air pollution," Int. J. Appl. Math. Comput. Sci., vol. 26, no. 2, pp. 467–478, 2016, doi: 10.1515/amcs-2016-0033.
[22] I. D. Borlea, R. E. Precup, F. Dragan, and A. B. Borlea, "Centroid update approach to K-means clustering," Adv. Electr. Comput. Eng., vol. 17, no. 4, pp. 3–10, 2017, doi: 10.4316/AECE.2017.04001.
[23] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," BMC Public Health, vol. 17, no. 1, pp. 1–19, 2017, doi: 10.1186/s12889-017-4914-3.
[24] Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris, "A comparative study for outlier detection techniques in data mi