EXPLORATORY DATA ANALYSIS AND

PREDICTION ON INDIA’S AIR QUALITY INDEX

USING MACHINE LEARNING TECHNIQUES

**ALEKYA RAMANI**

School of Computer Science Engineering

VELLORE INSTITUTE OF TECHNOLOGY

CHENNAI, INDIA

*alekya.ramani2020@vitstudent.ac.in*

**AASTHA BANSAL**

School of Computer Science Engineering

VELLORE INSTITUTE OF TECHNOLOGY

CHENNAI, INDIA

*aastha.bansal2020@vitstudent.ac.in*

**SANKAR KUMAR**

School of Computer Science Engineering

VELLORE INSTITUTE OF TECHNOLOGY

CHENNAI, INDIA

shankarmano17@gmail.com

**DR.TRILOKNATH PANDEY**

School of Computer Science Engineering

VELLORE INSTITUTE OF TECHNOLOGY

CHENNAI, INDIA

*triloknath.pandey@vit.ac.in*

***Abstract*—Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution.With this increasing air pollution,We Are in need of implementing models which will record information about concentrations of air pollutants(so2,no2,etc).The deposition of this harmful gasses in the air is affecting the quality of people’s lives, especially in urban areas. Lately, many researchers began to use the Big Data Analytics approach as there are environmental sensing networks and sensor data available.In this paper, machine learning techniques are used to predict the concentration of so2 in the environment. Sulfur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs.Models in time series are employed to predict the so2 readings in nearing years or months.**

***Keywords—air quality index, so2, no2, suspended particulate matter, respirable suspended particulate matter***

# Introduction

In developing countries like India, the rapid increase in population and economic upswing in cities have led to environmental problems such as air pollution, water pollution, noise pollution and many more. Air pollution has a direct impact on human health .There has been increased public awareness about the same in our country.Global warming, acid rains, increase in the number of asthma patients are some of the long-term consequences of air pollution. Precise air quality forecasting and thence producing proper control measures can reduce the effect of maximal pollution on the humans and biosphere as well. Accurate forecasting helps people plan ahead, decreasing the effects on health and the costs associated. If people are aware of variations in the quality of the air they breathe, the effect of pollutants on health as well as concentrations likely to cause adverse effects and actions to curtail pollution.

Similar to forecasting weather, there are models to predict levels of air pollution and air quality. There are many forecast models that require more complexity than weather forecast models. These models are mathematical simulations of how airborne pollutants disperse in the air. Machine learning algorithms can help in predicting the AQI. Linear regression, LASSO regression, ridge regression and Random Forest for regression were used to forecast the AQI.

Sometimes it is convenient for media platforms like news reporters and forecast magazines to print the required content in a simply understandable Language. Hence, to make this problem a classification problem, the AQI is calculated into various spans and are labeled as “Good”, “Poor”, “Moderate”, “Unhealthy”,”Very Unhealthy” and “Hazardous”.

Classification Algorithms such as Logistic Regression, Random Forest Classifier, KNN, Weighted KNN, Ridge Classification, AdaBoost Classifier and XGBoost Classifier are used on the dataset. Almost all of the models have performed extremely well, the results have been shown later.

II. Literature review

*Previous studies show the need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area [3][2]. The aim of this research paper is to investigate various air quality forecasting methods.*

*In recent years, there has been a lot of research done on the impact of air pollution on health [4]. Increases in mortality and hospital admissions from respiratory and cardiovascular illness have been linked to exposure to pollutants including ozone and airborne particulate matter.*

*The impact of variations in air quality on human health has been studied through short- and long-term epidemiological studies as well as sporadic air pollution episodes like the infamous London fog in 1952. The link between air pollution and higher mortality and hospital admissions is a recurring finding [6]..*

*Both short-term research that link daily changes in air pollution and health and long-term studies that have tracked cohorts of exposed people over time have discovered these impacts. Consequences have been observed at extremely low exposure levels, and it is uncertain whether particulate matter and ozone have a threshold concentration below which no adverse health effects are predicted [4].*

*There have been extensive studies conducted right from the 1900s after recognising the dire effects of the poor quality of air in the environment of an individual. Earlier studies were based on association with the direct relationship between the nature of a patient’s symptoms and the carbon pollutants from sources like gasoline and coal. From national consensus for fog control and smoke abatement in 1900-1970s, global AQI standards have been set and also consideration for the future generations were being put in place with amendments for the current state of air quality and also to prevent further degradation of the air quality[7][8]*

*The nature of air is impacted by multi-faceted elements including area, time, and unsure factors. Past investigates show to think about elements, for example, of air contaminants [1] like NO2, CO, Ground level O3, SO2, PM2.5(particulate matter with diameter of 2.5\*(10^-6) m), PM10, and meteorological data [3] like temperature, pressure, humidity, wind speed, wind direction.*

*The proposed algorithms in the current researches include ML techniques like Decision-Trees, Multiple Linear Regression (MLR), Multilayer Perceptron Artificial NN, Decision-Trees before ANN, random forest, back propagation and more [1][3][5].8].*

*Methods used for evaluating the predictive models (calculation of accuracies) were Root Mean Square Error, Normalised Root Mean Square Error, Mean Absolute Error, Symmetric Mean Absolute Percentage Error and Pearson correlation coefficient [1][5]. The results showed that compared to other methods, SLI-ESN (hybrid Scalable Link Interface-Echo state networks) performed better results [1].*

## It can be noted that most extensive research has been done in China which is leading this kind of work with 26 papers, followed by nations like Italy, Spain, USA, and more [1].

## 

## Given the high cost of further measures to reduce air pollution, and the many new findings which suggest that health effects can be seen at ever lower concentrations, the health effects of air pollution will need to receive much scientific and regulatory interest for years to come [4].

## therefore, we conclude that

## -Rather than utilizing straightforward AI strategies. Presently, the authors apply advanced and sophisticated techniques like gradient boost algorithms, random forest, Neural networks, back propagation [1][2][5].

## -China was the leading main nation as far as such investigations are considered [3].

## -Particulate matter with measurement equivalent to 2.5 micrometers was the fundamental expectation target [1].

## -Because of the number of deaths due to pollutants like SO2, PM10 and PM2.5, health effects from air pollution have been estimated to be higher than effects from a long list of other environmental factors [4][6].

## -The applications of air quality forecasting methods could be applied for air quality management purposes and protect public health.

##### III. BACKGROUND AND MOTIVATION

The introduction of dangerous or excessive amounts of specific compounds into the atmosphere is the main source of air pollution. These substances include gas, liquid droplets, and solid particles. Primary and secondary pollutants are the two types of air pollutants. Primary pollutants are those that are directly emitted into the atmosphere from their source. Primary air pollutants can come from both natural and man-made sources, such as burning fossil fuels, leaking gas from appliances, and volcanic eruptions and sandstorms. Sulfur dioxide (SO2), nitrogen oxides (NOx), particulate matter (PM), and carbon monoxide are examples of primary pollutants (CO). When primary pollutants interact chemically or physically, secondary pollutants are created in the atmosphere. Photochemical oxidants and secondary particulate matter are examples of secondary pollutants.

Criteria pollutants are the most abundant pollutants and they correspond to the most common health risks. These include SO2, lead, NO2, ground-level ozone (O3), and PM. It has been shown that there is a link between brief exposure to these pollutants and health problems like inflamed respiratory tract in healthy individuals, increased respiratory symptoms in asthmatics, trouble meeting high oxygen demands during exercise, and critical respiratory situations, especially in children and the elderly. Standards for acceptable air quality levels have been established by national organizations such as the EPA, EU, and many others. The levels of the air's criteria pollutants are shown by the air quality index (AQI).The highest AQI reading for each of the individual criterion pollutants makes up the total AQI. The health concerns connected to exposure to a specific air quality are also indicated by AQI levels. These medical symptoms may appear soon after being exposed to contaminated air or they may develop over time. Depending on the age and health status of the specific person being exposed, these symptoms may also differ. To prevent further rises in air pollution by on-demand pollution control systems or an emergency response, it is crucial that we have a system to foresee increases in air pollution levels. As a result, AQI would be easier to regulate to meet the demands of the population as a whole.

IV. RELATED WORK

In this research paper we have forecasted the air

quality of India by using algorithms to predict the air quality index(AQI) of a given area. Air qualityIndex is a standard measure to determine the quality of air.

Concentration of Gases such no2,co2, rspm, spm,

are recorded by the agencies . We have developed

a model to predict the air quality index based on the Historical Daily Ambient Air Quality Data released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP). This model is capable of successfully predicting the air quality index of a total county or any state or any bounded region provided with the historical data of pollutant concentration.This paper presents an integrated model using ArtificialNeural Networks and KNN to predict the level of air pollutants at various locations in India using past data available from Pollution Control Board. The proposed model is implemented and tested using python for and R.This system has used the Linear regression for prediction of AQ Index,Random Forest for prediction of AQI and AQI Range.This proposed system detects the levels of PM2.5 based on given atmospheric values. Logistic regression is used to detect whether a data sample is either polluted or not polluted.

V. DATA DESCRIPTION

The datasets used in this study were obtained from Github.This data is combined (across the years and states) and largely clean version of the Historical Daily Ambient Air Quality Data released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

The data consists of hourly concentrations of particulate matter of sizes less than or equal to 2.5 microns (PM2.5). They derive the Air Quality Index based on these values. The values are recorded using device located on the state.The data from Github are daily recordings of the concentrations of Sulfur dioxide (SO2), Nitrogen dioxide (NO2), PM2.5, and respirable suspended particulate matter (RSPM),*s*uspended particulate matter (SPM). The data is of relatively poor quality with a significant amount of missing values.The recordings are from all the states and union territory of India .The data is recorded monthly and the continuity of the time series representation of the data is desirable.

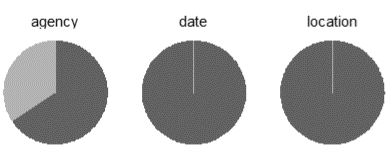
VI. DATA PREPROCESSING

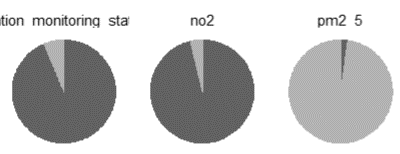
The dataset contains 56,64,646 values in total and 10,81,371 of them are missing. So proper pre-processing is vital for getting desired results. The attributes present in the dataset can be broadly divided into ‘categorical’ and ‘numerical’. ‘stn\_code’, ‘location\_monitoring\_station’ and ‘sampling\_date’ data do not contribute towards our results so they are removed. Looking through the dataset it can be observed that the last three rows majorly have missing values, so removing those rows would be better than imputing them. Day of the year and month have been extracted from the ‘date’ attribute and added to the record. Date attribute is removed.

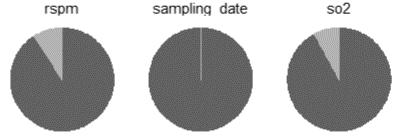
The categorical attributes - ‘agency’ has 34.30493%, ‘date’ has 0.001606455% and ‘type’ has 1.237659% of missing values. Missing values in agency have been replaced with ‘unknown’ as it is difficult to guess the agency. In the dataset date is generally repeated in a small localized range so the missing values have been replaced by the previously accessed date value. Similarly, missing values in type too have been replaced by previously accessed type values.

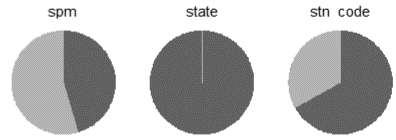
The numerical attributes – ‘so2’ has 7.951035%, ‘no2’ has 3.72537%, ‘spm’ has 54.4788%, ‘rspm’ has 9.230692% and ‘pm2\_5’ has 97.8625% of missing values. For so2, no2 and rspm the missing values have been replaced by state-wise mean values so that variance would not increase much and the results will not be skewed. Since spm and pm2\_5 majorly have missing values, they were imputed with ‘0’.

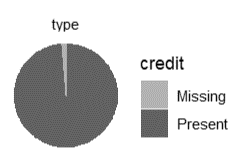
To find AQI values, SOi, NOi, SPMi and RSPMi have to be calculated. The required calculations have been done and for AQI, maximum among those four calculated values rows wise has been taken. The range for AQI has been calculated from the range given by the Government of India, Ministry of Environment, Forest and Climate Change.











Bihar, Delhi, Gujarat, Manipur and Uttaranchal have high pollution levels.

VII.ARCHITECTURE AND MODEL RESULTS:

Neural networks (ANNs) are computer programmes with biological inspiration that mimic how the human brain processes information.

In any neural network, a dense layer that is densely linked to the layer above it is one in which every neuron in the layer is coupled to every other neuron in the layer above.

In Dropout: A Simple Way to Prevent Neural Networks from Overfitting, published in 2014, is a regularization method for neural network models developed by Srivastava et al that explains about Dropout which is a training method in which certain neurons are disregarded at random.

The Callback lets us provide the performance indicator to track and the event that will halt training when it is triggered. When created, the EarlyStopping callback is set up using parameters.

Adam is the best among the adaptive optimizers in most of the cases. Good with sparse data: the adaptive learning rate is perfect for this type of datasets. Adam optimization algorithm is used because it can be used instead of the classical stochastic gradient descent procedure to update network weights iteratively based on training data.

**table 1.1 Regression it is Prediction of AQI**

| **REGRESSION** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **MODELS/METRICS** | **MAE** | **MSE** | **RMSE(train)** | **RMSE(test)** | **R2 SCORE(train)** | **R2 SCORE(test)** |
| **LINEAR REGRESSION** | **9.006** | **181.204** | **13.355** | **13.461** | **0.9853** | **0.985** |
| **RANDOM FOREST REGRESSOR** | **0.015** | **0.266** | **0.347** | **0.515** | **0.9999** | **0.9999** |
| **RIDGE REGRESSION** | **9.006** | **181.204** | **13.355** | **13.461** | **0.9853** | **0.985** |
| **LASSO REGRESSION** | **9.146** | **182.37** | **13.407** | **13.504** | **0.9852** | **0.9849** |
| **ANN** | **44.395** | **3573.9** | **59.91** | **59.78** | **0.999** | **0.999** |

**table 1.2 Prediction of AQI Range**

|  |  |  |  |
| --- | --- | --- | --- |
| **CLASSIFICATION** | | | |
| **MODELS/METRICS** | **ACCURACY(train)** | **ACCURACY(test)** | **KEPPA SCORE** |
| **LOGISTICS REGRESSION** | **0.8129** | **0.8125** | **0.7211** |
| **RANDOM FOREST REGRESSOR** | **0.999** | **0.99988** | **0.99983** |
| **KNN** | **0.98455** | **0.9738** | **0.96127** |
| **WEIGHTED KNN** | **0.98225** | **0.97038** | **0.95597** |
| **RIDGE REGRESSION** | **0.6571** | **0.6573** | **0.4275** |
| **ADABoost** | **0.753** | **0.753** | **0.6062** |
| **XGBoost** | **0.9997** | **0.9997** | **0.9996** |

VIII. CONCLUSION

More priority and resources have to be given to Bihar, Delhi, Gujarat, Manipur and Uttaranchal for repairing the AQI levels as they have high levels of pollution. The task of forecasting pollutant levels is inherently hard because of the volatile and dynamic nature of the data and its variability in space and time. However, the task of forecasting pollutant levels has been increasing in importance due to the effects of pollution on the population and the environment.In this work we have use algorithmic techniques like Linear Regression,Random Forest Regression,Logistic Regression,Random Forest Classifier,KNN,ANN(Artificial Neural Networks) for forecasting levels of pollutants like NO2, SO2, PM2.5 and Air Quality Index (AQI), using publicly available data for India.

## **Conflict of interest**

We do not have any conflict of interest with other authors.

References

1. Ditsuhi Iskandaryan , Francisco Ramos and Sergio Trilles,”Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review”,Institute of New Imaging Technologies (INIT), Universitat Jaume I, Spain,Received: 07 February 2020; Accepted: 25 March 2020; Published: 1 April 2020
2. Mrs. A. Gnana Soundari MTech, (PhD) Associate Professor,Mrs. J. Gnana Jeslin M.E, (PhD) Assistant Professor,Akshaya A.C’s,”INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING”,Jeppiaar Engineering College
3. Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Approaches”,Gang Xie,”Air Quality Prediction: Big Data and Machine Learning
4. Brunekreef B, Holgate ST. Air pollution and health. Lancet. 2002 Oct 19;360(9341):1233-42. doi: 10.1016/S0140-6736(02)11274-8. PMID: 12401268.
5. Jamal A, Nabizadeh Nodehi R. PREDICTING AIR QUALITY INDEX BASED ON METEOROLOGICAL DATA: A COMPARISON OF REGRESSION ANALYSIS, ARTIFICIAL NEURAL NETWORKS AND DECISION TREE. JAPH. 2017;2(1).
6. Marilena Kampa and Elias Castanas’,“Human health effects of air pollution”Laboratory of Experimental Endocrinology, University of Crete, School of Medicine, P.O. Box 2208, Heraklion, 71003, Greece. Received 4 June 2007, Accepted 10 June 2007, Available online 23 July 2007.
7. Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, Alexander Baklanov,Real-time air quality forecasting, part I: History, techniques, and current status, Atmospheric Environment,Volume 60,2012,Pages 632-655,ISSN 1352-2310
8. John Bachmann (2007) Will the Circle Be Unbroken: A History of the U.S. National Ambient Air Quality Standards, Journal of the Air & Waste Management Association, 57:6, 652-697, DOI: 10.3155/1047-3289.57.6.652