CHAPTER 23
A Machine Learning based Air Pollution Prediction in Smart Cities

Mr. S. MOHAN
Nehru Institute of Engineering and Technology, India

Mrs. R. SARANYA
Nehru Institute of Engineering and Technology, India

Mr. V. SATHEESWARAN
Nehru Institute of Technology, India

Mr. K. ARUNPATRICK
Nehru Institute of Technology, India

ABSTRACT

This paper presents the evaluation of air quality status performed during the period 2020 at three different sites (Industrial Area, Residential, Rural and other Areas) in the major cities of Tamilnadu. We attempted to investigate three major pollutants (PM10, NO2, and SO2) and their 24 hourly Concentrations were used for the calculation of the ambient air quality using IND-AQI procedure. It has observed that the calculated air quality values for SO2 and NO2 under „good" category. The calculated air quality values of PM10 for Particular areas (District like Chennai, Thoothukudi) are more than prescribed standard given by Central Pollution Control Board, New Delhi, India. The overall air quality was found to fall under the category „satisfactory" and „moderate" owing to PM10. In this Study Machines Language Techniques was performed to identify relationship between dependent variable (air quality) and independent variable (or) features (SO2, NO2, and PM10…, so on). We are also performing a category of Machines Language which is called Supervised Learning as we are training our Model with a given data set. In this exemplary study, we developed Predictive Models employing four Machines Language methods to identify air quality based on pollutants, district level and types of locations. Although the performances varied slightly Random forest achieved about 100% accuracy on predicting air quality and compared to Support Vector Machine (SVM) model that achieved about 99.96% predictive accuracy, KNN model that achieved about 99.52% predictive accuracy and Logistic Regression that achieved about 93.74% predictive accuracy. This study is a testament to the improving capabilities of Machines Learning Techniques in support for policy makers to take quick and broad decisions to improve air quality.

Keywords: *KNN, SVM, Machine Learning, PM10, NO2, and SO2*

INTRODUCTION

Air pollution is a problem for much of the developing world and is believed to kill more people worldwide than AIDS, Malaria, Breast Cancer or tuberculosis . India adopted the ambient air quality standard and began development of a nation-wide programme of ambient air quality monitoring known as National Air quality monitoring program(NAMP). The network consists of 683 operating stations covering three hundred cities/towns in 29 states and four union territories of the country. The monitoring

ISBN: 978-93-5782-920-5, www.aetsjournal.com "Handbook of research in Big data Analytics, Artificial intelligence and Machine learning"

of pollutants is carried out for 24 hr (4 hourly sampling for gaseous pollutants and 8 hourly sampling for particulate matter) with a frequency of twice a week, to have 104 observations in a year. The environmental regulatory authorities have prescribed the guidelines for the maximum permissible levels of various air pollutants, such as SO2(80 µg m-3),NO2(80 µg m-3),and PM10(100 µg m- 3),respectively. Urban air pollution in India had increased rapidly with the population growth, number of motor vehicles, use of fuels with poor environmental performance, badly mentioned transportation system, poor land use pattern, and above all ineffective environmental regulations. As for the health impact of air pollutants, air quality index (AQI) is an important indicator for general public to understand easily how bad or good the air quality is for their health and to assist in data interpretation for decision making processes related to pollution mitigation measures and environmental management. Basically the AQI is defined as an index or rating scale for reporting the daily combined effect of ambient air pollutants recorded at the monitoring site Machine learning is to predict the future from past data. Computer studying (ML) is a style of artificial intelligence (AI) that delivers computers the capability to gain knowledge of without being explicitly programmed. Machine finding out makes a speciality of the progress of pc applications that can alternate when exposed to new information and the basics of laptop studying, implementation of a easy laptop finding out algorithm utilising python. Process of coaching and prediction involves use of specialised algorithms. It feed the training data to an algorithm, and the algorithm uses this training knowledge to offer predictions on a brand new test information. Machine finding out can be roughly separated in to three classes. There are supervised learning, unsupervised finding out and reinforcement finding out. Supervised studying software is each given the input knowledge and the corresponding labeling to be trained data must be labeled with the aid of a person previously. Unsupervised learning isn't any labels It provided to the learning algorithm. This algorithm has to figure out the clustering of the input knowledge. Subsequently, Reinforcement learning dynamically interacts with its environment and it receives positive or bad suggestions to toughen its efficiency.

The necessity of healthy air has always been of great importance. As air is vital for all living beings on earth, it is our responsibility to keep the air clean. The rapid urbanization and industrialization have led the world into a new era of air pollution and is seen as a modern-day curse. Air pollution refers to the contamination of the air by excessive quantities of harmful substances. Most air pollution occurs from energy use and production, where emissions from traffic and industry are major contributors. Air pollution is a widespread problem due to its impact on both humans and the environment. Urban cities usually have the worst air pollution due to human activities [Kampa and Castanas (2008)]. Clear links between pollution and health effects have been revealed, which includes both short- and long-term consequences [Brauer et al. (2012)]. Associations with reduced lung function and increase in heart attack [Arden et al. (2002)], direct impact on people with asthma and other types of pneumonia [Guarnieri and Balmes (2014)] and once inhaled, a fine particular matter may hardly be self-purified by the immune system [Becker (2002)]. The overall effects of ambient air pollution on premature human mortality are a falling global trend, but in a smaller geographical area, the levels do not follow WHO‟s guidelines [WHO et al. (2006)]. Due to these severe problems, there are national requirements and objectives that each city must meet. Air quality has increasingly attracted attention from environment managers and citizens all over the world. New tools continue to emerge to raise air quality awareness worldwide.

Continuously improvements in air quality mapping are happening along with the advancements of smart cities and the amount of internet-of-things sensor devices. The increase in data produced contributes to further momentum in air pollution activity. A hot research topic is air pollution forecasting, the prediction of the atmospheric composition of pollutants for a given time and location. With an accurate air quality forecast, one can decide how to act due to air pollution health effects. On the national level, accurate forecasting contributes to planning and establishing procedures to reduce the severity of local pollution levels. With better knowledge at the individual level, one can choose the right choice for the cleanest routes for the commute, the best time for outdoor activity and other daily outdoor activities. Awareness

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

like this has the potential to create a cleaner environment and a healthier population. Accurate time series forecasting of air quality is a continuous research area, and much effort has been made by researchers to create models capable of fitting the underlying time series. Often, air quality prediction involves a noisy and limited amount of historical data. Furthermore, the prediction of a single observation usually depends on many events that rely on each other. The models are then forced to include specially adapted techniques to comply with the erroneous or lack of data.

These complex problems make it hard to generalize the solution to be transferable to other locations. Besides, the air changes rapidly in short time frames, with hourly data more uncertain compared with monthly and yearly trends and seasonality. The lack of and poor quality of data, a low spatial resolution of data points, and the cost of high-quality sensors add up to the list among other obstacles. Figure 1.1 presents the observations of particular matter smaller than 2.5 microns (PM2.5) in the period of the end of 2017 to Mai 2019. The graph shows the typical pollution trends in Trondheim, of rapid changes in pollution levels for the winter months, while the summer includes a relatively good level. These trends and patterns are typical characteristics of the air quality in cities in Norway and Scandinavia. Consequently, the challenge of these time series is then the prediction of the sudden changes in harmful pollution.

LITERATURE SURVEY

Verma, Ishan, Rahul Ahuja, HardikMeisheri, andLipikaDey. ”Air pollutant severity rediction using Bi- directional LSTM Network.” In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI),pp. 651-654. IEEE, 2018.

This author described the benefits of the Bidirectional Long -Short Memory [BiLSTM] method to forecast the severity of air pollution. The proposed technique achieved better prediction which models the long term, short term, and critical consequence of PM2.5 severity levels. In the proposed method prediction is made at 6h, 12h, 24h. The results obtained for 12h is consistent, but the result obtained for 6h, and 24h are not consistent.

Figures Zhang, Chao, Baoxian Liu, Junchi Yan, Jinghai Yan, Lingjun Li, Dawei Zhang, XiaoguangRui, and RongfangBie. ”Hybrid Measurement of Air Quality as a 5 Fig. 8. RH w.r.t tin oxide Fig. 9. RH w.r.t C6H6 Mobile Service: An Image Based Approach.” In 2017 IEEE International Conference on Web Services (ICWS), pp. 853-856. IEEE,2017.

This author proposed web service methodology to predict air quality. They provided service to the mobile device, the user to send photos ofair pollution. The proposed method includes 2 modules a) GPS location data to retrieve the

assessment of the quality of the air from nearby air quality stations. b) they have applied dictionary learning and convolution neural network on the photos uploaded by the user to predict the air quality. The proposed methodology has less error rate compared to other algorithms such as PAPLE, DL, PCALL but this method has a disadvantage in learning stability due to this the results are less accurate.

Yang, Ruijun, Feng Yan, and Nan Zhao. ”Urban air quality based on Bayesian network.” In 2017 IEEE 9th Fig. 10. RH w.r.t NO Fig. 11. RH w.r.t NO2 International Conference on Communication Softwareand Networks (ICCSN), pp. 1003-1006. IEEE,2017.

This author used the Bias network to find out the air quality and formed DAG from the data set of the town called as shanghai. The dataset is dived for the training and testing model. The disadvantage of this approach is they have not considered geographical and social environment characteristics, so the results may vary based on these factors.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Ayele, TemeseganWalelign, and RutvikMehta.”Air pollution monitoring and prediction using IoT.” In 2018 Second International Conference on Inventive Communication 6Fig. 12. RH w.r.t Temperature Fig. 13. RH w.r.t CO and Computational Technologies (ICICCT), pp. 1741-1745. IEEE,2018.

This author proposed an IoT based technique to obtain air quality data set. They have used Long Short-term Memory[LSTM] technique in-order to predict the air quality the proposed technique achieved better accuracy by reducing the time taken to train the model. But still, the accuracy can be improved by compared other techniques such as the Random forest method

Djebbri, Nadjet, and MouniraRouainia. ”Artificial neural networksbased air pollution monitoring inindustrial sites.” In 2017 International Conference on Engineering and Technology (ICET), pp. 1-5. IEEE,2017.

This author proposed artificial based Regressive model which is nonlinear to predict 2 major air pollutants 2 such as carbon monoxide and nitrogen oxides. They have considered the variables such as the speed of the air, air direction, temperature, and moisture and the toxic elements from the industrial site such as Skikda. They have used RMSE and MAE to determine the performance, but this method considered only 2 pollutants such as NO and CO the other major pollutants such as sulfur dioxide, PM2.5, PM10 are not considered.

D. Van Le, C. K. Tham, Machine Learning (ML)-Based Air Quality Monitoring Using Vehicular Sensor Networks. InParallel and Distributed Systems (ICPADS), 2017 IEEE 23rd International Conference on 2017 Dec 15 (pp. 65–72). IEEE.

This author proposes a machine learning based Air quality Monitoring system which aims at reduce the sensing cost and communication by allowing vehicles to process the collected data in a distributed fashion. It was proposed to assign and sense locations to vehicles such that the successful measurement probability of all sensing sub-areas is maximized while the vehicles can learn models with prediction accuracy.

S. Kumar, A. Jasuja, Air quality monitoring system based on IoT using Raspberry Pi. In Computing, Communication and Automation (ICCCA), 2017 International Conference on 2017 May 5 (pp. 1341– 1346). IEEE.

This author proposes a real time AQI Monitoring System using various parameters such as CO, CO2, Humidity, PM 2.5, Temperature and air pressure. The system is tested in Delhi and the measurements are compared with the data provided by the local environmental control authority. Pollution Monitoring Sensor, Arduino Uno, Raspberry pi are used for monitoring the air quality, with accurate, affordable and easy to use. Furthermore, durable pollution arrangements can be detected and assured network between the air pollutants can be found.

K.S. Goverdhan Rathla,T. Sankarappa*, J.S. Ashwajeet and R. Ramanna- Air Quality Analysis using Machine Learning Algorithm

Ambient Air Quality (AQI) of Belagavi city and its surroundings was monitored using a set of dust samplers, viz., APM-460, APM-550 and APM-433. The pollutants considered for the assessment of air quality were sulphur dioxide (SO2), nitrogen dioxide (NO2), ammonia (NH3) and particulate matter (PM10 and PM2.5) which gave a measure of pollution content in the air. The sampling sites have been monitored for 24 hours on an 8 hourly basis in three different seasons in 2013 and 2014. Air quality was quantified in terms of concentrations of PM10, PM2.5, NO2, SO2 and NH3. The results revealed that locations near Bus station and Macchae industrial area lie in the range of Moderate Air Pollution and Autonagar, Sadashivnagar and Kangralli village lie in the range of Light Air Pollution. Annual Air Quality index of presently selected sites of Belagavi has been determined to assess the degree of atmospheric pollution.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Ioannis N. Athanasiadis, Vassilis G. Kaburlasos, Pericles A. Mitkas and Vassilios Petridis: Predicting and Analyzing Air Quality using Machine Learning: Comprehensive Model

Air quality is typically assessed based on either expert meteorologist knowledge or on sophisticated “first principles” mathematical models. Air Quality Operational Centers have been established worldwide in areas with (potential) air pollution problems. These centers monitor critical atmospheric variables and they publish regularly their analysis results [KA99]. Currently, real- time decisions are made by human experts, whereas mathematical models are used for off-line study and understanding of the atmospheric phenomena involved

Andrzej ŻYROMSKI, Małgorzata BINIAK-PIERÓG, Ewa BURSZTA ADAMIAK, Zenon ZAMIAR: Review on Data Mining Techniques for Prediction of Air Quality

The evaluation of the relation between meteorological elements and air pollutants‟ concentrations. The analysis includes daily concentrations of pollutants and variation of meteorological elements such as wind speed, air temperature and relative humidity, precipitation and total radiation at four monitoring stations located in the province of Lower Silesia in individual months of the winter half-year (November–April, according to hydrological year classification) of 2005– 2009. Data on air quality and meteorological elements came from the results of research conducted in the automatic net of air pollution monitoring conducted in the range of the State Environment Monitoring. The effect of meteorological elements on analyzed pollutant concentration was determined using the correlation and regression analysis at significance level $\alpha < 0.05$. The occurrence of maximum concentration of NO, NO2, NOX, SO2 and PM10 occurred in the coldest months during winter season (January, February and December) confirmed the strong influence of “low emission” on air quality. Among the meteorological factors assessed wind speed was most often selected component in step wise regression procedure, then air temperature, less air relative humidity and solar radiation. In the case of a larger number of variables describing the pollution in the atmosphere, in all analyzed winter seasons the most common set of meteorological elements were wind speed and air temperature.

EXISTING SYSTEM

Gaps in the Literature Air quality prediction has been widely researched using popular machine learning techniques. However, several gaps in the literature have been discovered. The most evident is the lack of a strict evaluation framework. The researchers use different problem definitions and evaluation methods to showcase their results. The problems include a large area of combinations: univariate or multivariate, fine-grained or single target predictions, and a window horizon of a few hours or a range of multiple days. Besides, the datasets used in the literature is tied up to the research location. The cities of interest each introduce a new set of data with different distribution and variables. The data is characterized by the cities unique geographical factors, climate, and the city magnitude. The datasets from a town might be different from another or challenging to obtain. These limitations make the promising solutions from the literature challenging to reproduce for other locations. However, this is not the focus of this thesis. The impact of events and human mobility data on air quality patterns has not yet been studied, as far as we are aware. This kind of data might be of challenge to acquire or includes inaccurate information, and thus not applicable for experiments. The city population tends to gather around more significant events, and therefore, the air quality could have a correlation with this movement of population density. Despite this interesting direction, this was not followed through, due to that, we were not able to acquire the necessary data in time. The literature includes comprehensive experiments of meteorological, temporal, and spatial techniques. These features are further used to highlight temporal and spatial relations by using feature engineering. These spatial relations are described with neighboring air quality measurements. The temporal relations are represented with historical data and with including information of the timestamp of the measure. Another approach is to include statistical calculations of the time series to complement the features. This approach is less seen from the research and is believed to add even more hidden relations of the complexity in air quality, and is studied in this thesis.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

OVERALL ARCHITECTURE

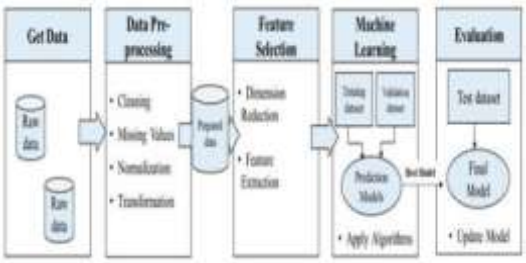


Figure 4.1 Overall Architecture of Proposed System

The basic aim of our project is to apply Machine Learning techniques based classification and regression functions to enumerate the influence of pollutants on the air quality and to predict the air quality index in the study region using primary pollutants (SO2, NO2 and PM10) as the estimators. Accordingly, in this project we used to predict air quality by using Multiple Classifier in Machine Learning methods (Decision tree, KNN, SVM, Random forest, Logistic Regression, Naive Bayes) for classification and regression.

LIST OF MODULES

1. Build Data Set
2. Data Pre-processing
3. Feature Selection
4. Train the Model
5. Test the Model
6. Data Analysis
7. Prediction

BUILD DATA SET

Tamil Nadu is the eleventh-largest state in India by area and the sixth-most populous. Under NAMP, Ambient Air Quality is being monitored by CPCB in association with Tamil Nadu Pollution Control Board (TNPCB) in 28 locations covering cities, major towns and major industrial areas viz. Chennai, Salem, Coimbatore, Madurai, Trichy, Cuddalore, Mettur, and Thoothukudi. All these stations are manual operated stations. The ambient air samples are collected through high volume samplers by running 24 hours and twice a week. Thus in each stations, not less than 108 samplings are done in a year. PM10, SO2 and NO2 are monitored. Out of these 28 stations, 10 stations were selected to calculate historical AQI so as to know the air quality of the cities and towns. With this, the Air quality index ranges will be generated and are organized in a CSV (Comma Separated Values) file.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

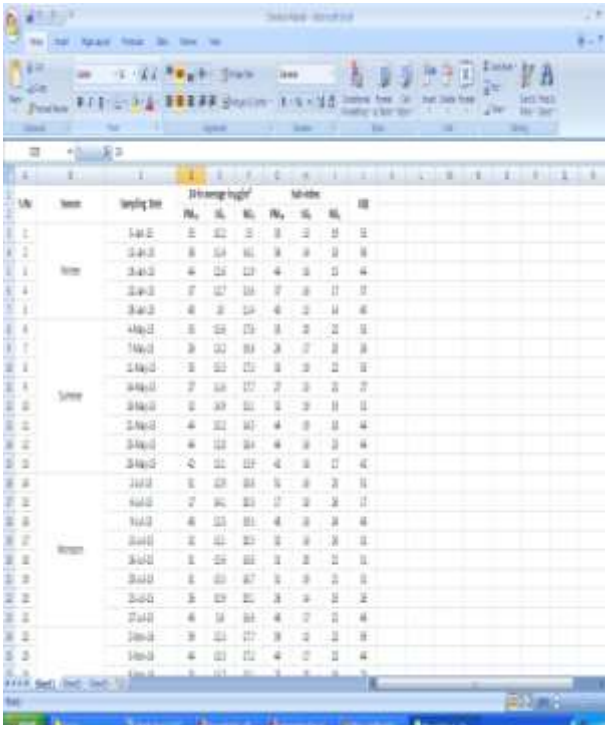


Figure 4.2 Air Quality Parameters- Chennai

AQI Category (Range)	PM10 24-hr	PM2.5 24-hr	NO2 24-hr	O3	CO	SO2 24-hr	NH3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5-1.0
Moderately Polluted (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200	2.1-3.0
Very Poor (301- 400)	351-430	121-250	281-400	209-748*	17-34	801-1600	1200-1800	3.1-3.5
Severe(401-500)	430 +	250 +	400 +	748+*	34 +	1600+	1800+	3.5+

Table-4 Air Quality Index scale and its categories

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

AQI	Associated Health Impacts
Good (0-50)	Minial Impact
Satisfactory(51-100)	May cause minor breathing discomfort to sensitive people
Moderately polluted (101-200)	May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children and older adults
Poor (201-300)	May cause breathing discomfort to the people on prolonged exposure and discomfort to people with heart disease
Very Poor (301- 400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
Severe (401-500)	May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

Table-5 Health Statements for IND-AQI Categories

DATA PRE-PROCESSING

The data we get from different sources may contain inconsistent data, missing values and repeated data. To get proper prediction result, the dataset must be cleaned, missing values must be taken care of either by deleting or by filling with mean values or some other method. Also redundant data must be removed or eliminated so as to avoid biasing of the results. Some dataset may have some outlier or extreme values which also have to be removed to get good prediction accuracy. Classification and clustering algorithms and other data mining methods will work well only if all this pre- processing is done on the data.

Data preprocessing in Machine Learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In simple words, data preprocessing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format. When it comes to creating a Machine Learning model, data preprocessing is the first step marking the initiation of the process. Typically, real-world data is incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends. This is where data preprocessing enters the scenario – it helps to clean, format, and organize the raw data, thereby making it ready-to-go for Machine Learning models.

Identifying and handling the missing values

In data preprocessing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data.

Needless to say, this will hamper your ML project.

Basically, there are two ways to handle missing data:

Deleting a particular row – In this method, you remove a specific row that has a null value for a feature or a particular column where more than 75% of the values are missing. However, this method is not 100% efficient, and it is recommended that you use it only when the dataset has adequate samples. You must ensure that after deleting the data, there remains no addition of bias.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Calculating the mean – This method is useful for features having numeric data like age, salary, year, etc. Here, you can calculate the mean, median, or mode of a particular feature or column or row that contains a missing value and replace the result for the missing value. This method can add variance to the dataset, and any loss of data can be efficiently negated. Hence, it yields better results compared to the first method (omission of rows/columns). Another way of approximation is through the deviation of neighbouring values. However, this works best for linear data.

Duplicate values: A dataset may include

the process of dealing with duplicates. In most cases, the duplicates are removed so as to not give that particular data object an advantage or bias, when running machine learning algorithms.

ENCODING THE CATEGORICAL DATA

Categorical data refers to the information that has specific categories within the dataset. In the dataset cited above, there are two categorical variables – country and purchased.Machine Learning models are primarily based on mathematical equations. Thus, you can intuitively understand that keeping the categorical data in the equation will cause certain issues since you would only need numbers in the equations. The Air quality Data initially obtained must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format.

In this study, both the single pollutant and combined pollutants based AQIs were calculated for all the data points, single pollutant based AQI were calculated using

IND-AQI method

data objects which are duplicates of one another. It may happen when say the same person submits a form more than once.

Table 1. IND-Air Quality Index scale and its categories

*One hourly monitoring(for mathematical calculation only)

CALCULATION OF AQI

The CPCB has given guidelines on calculating AQI as follows [2]:

1. The Sub-indices for individual pollutants at a monitoring location are calculated using its 24-hourly average concentration value (8-hourly in case of CO and O3) and health breakpoint concentration range. The worst sub-index is the AQI for that location.
2. All the eight pollutants may not be monitored at all the locations. Overall AQI is calculated only if data are available for minimum three pollutants out of which one should necessarily be either PM2.5 or PM10. Else, data are considered insufficient for calculating AQI. Similarly, a minimum of 16 hours‟ data is considered necessary for calculating sub-index
3. The sub-indices for monitored pollutants are calculated and disseminated, even if data are inadequate for determining AQI. The Individual pollutant-wise sub- index will provide air quality status for that pollutant.
4. The web-based system is designed to provide AQI on real time basis. It is an automated system that captures data from continuous monitoring stations without human intervention,
5. and displays AQI based on running average values (e.g. AQI at 6 A.M on a day will incorporate data from 6 A.M on previous day to the current day).
6. For manual monitoring stations, an AQI calculator is developed wherein data can be fed manually to get AQI value.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

The data were then partitioned into two subsets; training and test set, In the present study, the complete data set (433 samples x 16 variables) was partitioned a training (433 samples x 16 variables) and test (337 samples x 16 variables)set. Thus the training and test sets comprised of 100% and 65% samples, respectively. Since all the independent variables (Air pollutants Concentration in (µg m-3), type of locations) in data exhibited significant correlations with the dependent variables, all of them were considered for Classification and regression.

DATA CLEANING

Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, deduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers believed to be reliable. Unusual amounts above or below pre-determined thresholds may also be reviewed. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. Textual data spellcheckers can be used to lessen the amount of mistyped words, but it is harder to tell if the words themselves are correct.

EXPLORATORY DATA ANALYSIS

Once the data is cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data. The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative in nature. Descriptive statistics such as the average or median may be generated to help understand the data. Data visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

AIR QUALITY CALCULATED USING AQI

Air quality index is a 100 point scale that summarizes results from a total of nine different measurements when complete: Year, Month, Season, Location, SO2, NO2, PM10, and Sub Index PM10. And each parameter Q-Values changes depends upon the their values.

$AQI = \Sigma AnWn / \Sigma Wn$

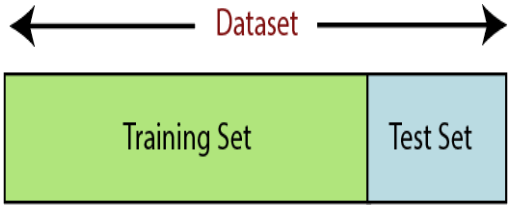
An = quality rating of nth Air quality parameter.

Wn = unit weight of of nth air quality parameter.

TRAIN THE MODEL

Training set denotes the subset of a dataset that is used for training the machine learning model. Here, you are already aware of the output. A test set, on the other hand, is the subset of the dataset that is used for testing the machine learning model. The ML model uses the test set to predict outcomes.

Usually, the dataset is split into 70:30 ratios or 80:20 ratios. This means that you either take 70% or 80% of the data for training the model while leaving out the rest 30% or 20%. The splitting process varies according to the shape and size of the dataset in question.



ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Figure 4.3 Train set and Test Set representation

x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.2, random_state=0)

Here, the first line splits the arrays of the dataset into random train and test subsets. The second line of code includes four variables:

- x_train – features for the training data
- x_test – features for the test data
- y_train – dependent variables for training data
- y_test – independent variable for testing data

Thus, the train_test_split() function includes four parameters, the first two of which are for arrays of data. The test_size function specifies the size of the test set. The test_size maybe .5, .3, or .2 – this specifies the dividing ratio between the training and test sets. The last parameter, "random_state" sets seed for a random generator so that the output is always the same. In machine learning, training data is the data you use to train a machine learning algorithm or model. Training data requires some human involvement to analyse or process the data for machine learning use. How people are involved depends on the type of machine learning algorithms you are using and the type of problem that they are intended to solve.

- With supervised learning, people are involved in choosing the data features to be used for the model. Training data must be labelled - that is, enriched or annotated - to teach the machine how to recognize the outcomes your model is designed to detect.
- Unsupervised learning uses unlabelled data to find patterns in

the data, such as inferences or clustering of data points. There are hybrid machine learning models that allow you to use a combination of supervised and unsupervised learning.

Figure 4.4 Training Model

TEST MODEL

This part of the dataset is used to test our model hypothesis. It is left untouched and unseen until the model and hyper parameters are decided, and only after that the model is applied on the test data to get an accurate measure of how it would perform when deployed on real- world data.

AIR QUALITY DATA ANALYSIS

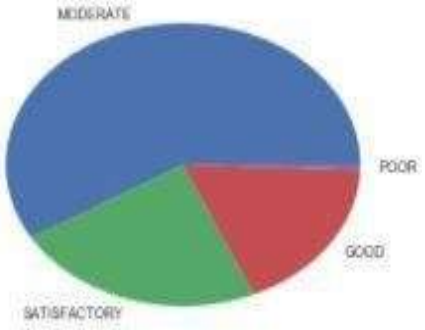
ISBN: 978-93-5782-920-5, www.aetsjournal.com "Handbook of research in Big data Analytics, Artificial intelligence and Machine learning"

The air quality data in the CSV file is used to perform analysis using machine learning algorithm. The analysis performed with the air quality data is the end result of the paper. There are 5 major analysis performed with the scrapped results.

RESULT AND DISCUSSION

The values of monthly PM10, SO2, NO2 concentrations measured in the major cities of Tamil Nadu, during the study period (2020-2021) The moderate level has been noted in Tamil Nadu during period 2007-2013 and satisfactory level has been noted during period (2020-2021) are shown table(5,6) respectively, where the high concentrations may be even more problematic in terms of human due to the fact that most of the particles in these locations are derived from anthropogenic sources. Their concentrations tended to increase for locations of various cities in Tamil Nadu, where they achieve moderate level during 2020 to 2021. On the contrary the concentrations significantly decrease in Chennai-Adayar (table5). The higher PM10 concentrations might be due to suspension of road dust, soil dust and vehicles and not good condition of the road. Table (7) presents a summary of the monthly averaged concentrations for oxides of nitrogen NO2, SO2, PM10 monitored during the whole period of the study. It can be seen that concentration levels are well below the prescribed IND- Air quality indices. Air pollution in Tamil Nadu was under permissible level. Level of PM10 in all stations in some locations found beyond the permissible limit (Satisfactory to Moderate level) but SO2 and NO2 were below the permissible limit at all the stations.

```
MODERATE        197
GOOD             77
SATISFACTORY     60
POOR              2
Name: AQI, dtype: int64
```



AIR QUALITY CLASSIFICATION

```
In [13]: aqchtr.describe(include=['object'])
```

Out[13]:

	Month	Season	Location	AQI
count	432	432	432	432
unique	12	3	4	4
top	october	summer	T.nagar	MODERATE
freq	36	144	108	214

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Fig-9 Unique values in analysis data

The values of monthly PM10, SO2, NO2 concentrations measured in the major location of Chennai, during the study period (2020-2021) The moderate level has been noted in Chennai during period 2020-2021 and satisfactory level has been noted during period (2020-2021) are shown table(3,4) respectively, where the high concentrations may be even more problematic in terms of human due to the fact that most of the particles in these locations are derived from anthropogenic sources. Their concentrations tended to increase for locations of T.nagar, Anna nagar and kilpauk, where they achieve moderate level during 2021 to 2020. On the contrary the concentrations significantly decrease in adayar(table3). The higher PM10 concentrations might be due to suspension of road dust, soil dust and vehicles and not good condition of the road. Table (5) presents a summary of the monthly averaged concentrations for oxides of nitrogen NO2, SO2, PM10 monitored during the whole period of the study. It can be seen that concentration levels are well below the prescribed IND- Air quality indices. Air pollution in Chennai area was under permissible level. Level of PM10 in all stations in some locations found beyond the permissible limit (Satisfactory to Moderate level) but SO2 and NO2 were below the permissible limit at all the stations.

OVERALL MONTH WISE PERFORMANCE OF AIR QUALITY INDICES STUDY AREAS

The air quality categories of air pollutants (PM10,SO2,NO2) of four different location based on the AQI, the Chennai area was categorized as (Adayar at residential area, Annanagar at residential area, T.nagar at commercial(traffic inter-section) and kilpauk at commercial (traffic inter- section)on type of location are presented in table 3. AQI values in this study were calculated by using the Concentration of PM10,SO2 and NO2 (by using standard formula mention in data processing). Air pollution in chennai different location in city was under control level. Level of PM10 in all four location satisfactory to moderate level. Levels of SO2, NO2 were below the permissible limit at all locations.

AIR QUALITY ANALYSIS FOR ALL DISTRICT

Accuracy is the overall number of the correct predictions fractionated by the whole number of predictions created for a dataset. It can inform us immediately if a model is trained correctly and by which method it may perform in general. Nevertheless, it does not give detailed information concerning its application to the issue. Precision, called PPV, is a satisfactory measure to determination, whereas the false positives cost is high. Recall is the model metric used to select the best model when there is an elevated cost linked with false negative. Recall helps while the false negatives‟ cost is high. F1-score is required when you desire to seek symmetry between both precision and recall. It is a general measure of the accuracy of the model. It combines precision and recall. A good F1-score is explained by having low false positives and also low false negatives.

True Positives (TP):- These are the correctly predicted positive values, which means that the value of the actual class is yes and the value of the predicted class is also yes.

True Negatives (TN):- These are the correctly predicted negative values, which means that the value of the actual class is no and value of the predicted class is also no.

False positives and false negatives, these values occur when our actual class contradicts with the predicted class.

False Positives (FP):- When actual class is no and predicted class is yes.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

False Negatives (FN):- When actual class is yes but predicted class in no.

Where,

- True positive (TP) = correctly identified
- False positive (FP) = incorrectly identified
- True negative (TN) = correctly rejected
- False negative (FN) = incorrectly rejected

Precision

Precision means to determine the number of positive class predictions

that actually belong to the positive class.

Precision = TP/TP+FP

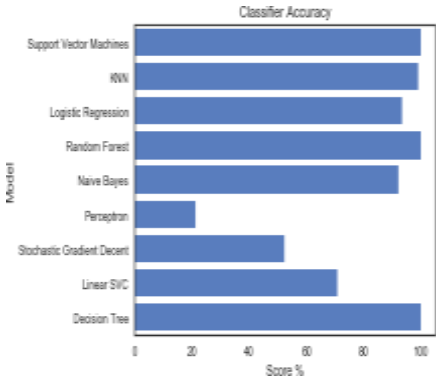


Figure 6.3

Predicted air quality using Machine Learning Approaches. Recall

Recall means to determine the number of positive class predictions made out of all positive samples in the dataset.

Recall = TP/TP+FN

Multiple Classifier	Score

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Support Vector Machines	100.00
Random Forest	100.00
Decision Tree	100.00
KNN	99.07
Logistic Regression	93.29
Naïve Bayes	92.13
Linear SVC	70.83
Stochastic Gradient	52.31

Table 7. Contribution of the predictor variables in the various Machine Learning models for prediction of air quality

The air quality categories of air pollutants (PM10,SO2,NO2) of major cities of Tamil Nadu based on the AQI, few cities of Tamil Nadu was categorized as Chennai-Adayar at residential area, Coimbatore at mixed zone, Chennai- T.nagar at commercial area and Madurai at mixed zone, Trichy at traffic zone on type of location are presented in table 5. AQI values in this project were calculated by using the Concentration of PM10, SO2 and NO2 (by using standard formula mention in data processing). Air pollution in many cities was under control level. Level of PM10 in all location satisfactory to moderate level. Levels of SO2, NO2 were below the permissible limit at all locations.

CONCLUSION

AQI of cities and towns in Tamil Nadu reveals that PM10 is the main contributor for higher value of index. SO2 and NO2 are well within the NAAQ standards for 24 hours. The higher value of PM10 is mainly due to vehicular pollution. Vehicular emissions are of particular concern because these are ground level sources and thus have the maximum impact on general population. Also, vehicles contribute significantly to the total air pollution load in many urban areas. It is to be noted that AQI system is based on maximum operator function by selecting the maximum of sub-indices of various pollutants as overall AQI. Ideally, eight parameters (i.e.,) PM10, PM2.5, NO2, SO2, CO, O3, NH3 and Pb having short-term standards should be considered for near real-time dissemination of AQI.

The regulating agencies should establish source-receptor relationships in terms of impact of emissions on air quality. Adopting comprehensive policies in an integrated manner and addressing the root causes rather than focusing on issues in isolation and seeking remedies is the key to managing air quality in urban areas. In case AQI category is severe or very poor, necessary steps need to be taken by further regulating the emissions which are causing maximum impact to ambient air quality. Specific actions, such as (i) strict vigilance and no-tolerance to visible polluting vehicles, industries, open burning, construction activities etc.; (ii) regulating traffic; and (iii) identifying sources contributing significantly to rising air quality levels and actions for reducing emissions from such sources are to be taken. In the cities well-constructed clean roads, flyovers, cleaner transport fuel will reduce the ambient air pollution level. AQI is an initiative intended to enhance public awareness and involvement in efforts to improve air quality. People can contribute by maintaining vehicles properly (e.g. get PUC checks, replace car air filter, maintain right tires pressure), following lane discipline & speed limits, avoiding prolong idling and turning off engines at red traffic signals.

REFERENCES

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Urban population (% of total population). https:// data. world bank. org/ indic ator/ SP. URB. TOTL. IN. ZS Accessed 20 Oct 2021.

Department of Economic and Social Affairs: Urban Population Change; 2018. https:// www. un. org/ devel opment/ desa/ en/ news/ popul ation/ 2018- revis ion- of- world- urban izati on- prosp ects. html. Accessed 20 Oct 2021.

Nada Osseiran, Christian Lindmeier: 9 out of 10 people worldwide breathe polluted air, but more countries are taking action; 2018. https:// www. who. int/ news/ item/ 02- 05- 2018-9- out- of- 10- people- world wide- breat he- pollu tedair-mbut- more- count ries- are- taking- action Accessed 20 July 2021.

Ailshire JA, Crimmins EM. Fine particulate matter air pollution and cognitive function among older US adults. Am J Epidemiol 2014;180(4):359–66. https:// doi. org/ 10. 1093/ aje/ kwu155. https:// acade mic. oup. com/ aje/ artic le- pdf/ 180/4/ 359/ 86408 02/ kwu155. pdf.

Pöschl U. Atmospheric aerosols: composition, transformation, climate and health effects. Angewandte Chemie Int Ed. 2005;44(46):7520–40. https:// doi. org/ 10. 1002/ anie. 20050 1122.

Du Y, Xu X, Chu M, Guo Y, Wang J. Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence. J Thoracic Dis. 2016;8(1):8.

Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, Balakrishnan K, Brunekreef B, Dandona L, Dandona R, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. Lancet. 2017;389(10082):1907–18.

Bu X, Xie Z, Liu J, Wei L, Wang X, Chen M, Ren H. Global pm2.5-attributable health burden from,. to 2017: estimates from the global burden of disease study 2017. EnvironRes. 1990;2021(197):11112

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”