Original Article

Urban Air Quality Analysis And Aqi Prediction Using Improved Knn Classifier

Krishna Chaitanya Atmakuri[1], K V Prasad[2]

[1]Research Scholar, [2] Associate Professor, Department of Computer Science and Engineering
[1,2] Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India
Email: [1]chaituit2004@gmail.com, [2]prasad_kz@yahoo.co.in
DOI: 10.47750/pnr.2022.13.S09.899

Abstract

Air pollution monitoring is a vexing problem in today life. Air pollution is one of the major variables that has the potential to negatively impact the quality of life of all living things in the environment. With the rapid development of various industries and motorised transportation, large amounts of harmful substances such as soot, sulphur dioxides, nitrogen oxides, particulate matter, humidity and hydrocarbons are released into the atmosphere, lasting for long periods of time and in concentrations that exceed tolerable environmental limits. So there is need to monitor the Air Quality by using the data. Machine learning techniques are used in this work to predict air AQI. An Improved KNN classifier is implemented to predict the AQI Several classifiers such as KNN, Logistic Regression, and Decision Tree algorithm are Compared used to analyse the findings depending on the given data in Python. Classifiers are compared based on how accurate their findings are. The best classifier is utilised for prediction when compared to the other models [8].

Keywords: Air Pollution, Machine Learning, Python, Classifiers, Prediction.

INTRODUCTION

Air contamination is quite possibly the most genuine natural issue defying current development. In most cases, it is a result of human activities such as transportation, mining, development, and contemporary labour. Air pollution is defined as the introduction of pollutants into the atmosphere as poisons that pose a threat to human health and the environment as a whole. WHO estimates that air pollution kills about seven million people throughout the globe annually.

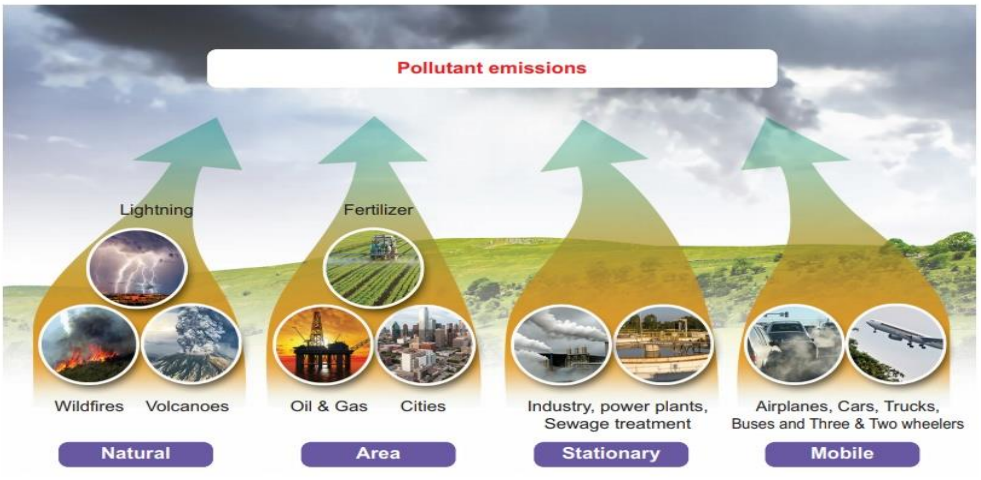

Figure 1: Air Pollutant Sources [7]

The Figure 1 depicts the various Air pollution sources. There are four types of sources like Natural, Area, Stationary, and Mobile. The Natural sources can be Lightning, Wildfires, and Volcanoes. The Area sources can be Fertilizer, Oil & Gas, and Cities. The Stationary sources can be Industry, Power plants, and Sewage treatment. The Mobile sources can be Airplanes, Cars, Trucks, Buses and Three & Two wheelers. All these sources conclude one thing i.e. emission of pollutants. So, these are the pollution sources that can occur either natural or through man made.

The WHO's pollution guideline levels are now exceeded by nine out of ten individuals, with the majority living in poor and middle-income nations. However, wildfires and volcanic eruptions contribute to air pollution, as can natural processes. Air pollution may be gaseous or solid, and both types are harmful to the environment (as suspended particulate matter in the air). Short-term health side effects of air pollution are just as serious as long-term health side effects of air pollution. Long-term side effects include lung cancer, liver damage, kidney damage, brain damage, and respiratory problems. Short-term, transitory effects include irritation of the nose, throat, eyes, and skin. Pollutants in the air may be:

1) Carbon monoxide is produced when fossil fuels, diesel, and gasoline are burned. Vehicles emit this gas, which is invisible and unnoticed. It has a negative impact on the lungs of humans, resulting in significant breathing difficulties.
2) Toxic air pollutants are produced by the combustion of fossil fuels. These are the most common causes of cancer and birth abnormalities.
3) Ozone (O3) is a poison that is surrounded by unstable natural chemicals when exposed to sunlight. It causes breathing difficulties, asthma, and a reduction in lung function.
4) Nitrogen Dioxide (NO2) is a pollutant produced by the combustion of fuels such as gasoline, industrial boilers, and wood, which causes lung ailments.
5) This sulphur and oxygen-based gaseous air pollution is known as sulphur dioxide (SO2). SO2 is created during the combustion of sulfide-containing fuels such as coal, oil, or diesel.

Many different types of models have been developed in order to forecast daily air quality. Some of them are:

The paper [1] tells about how to work with machine learning models. The proposed work is also using the machine learning algorithms to analyze, predict the data and also to know how to utilize the data such that the model is making correct predictions is learnt through this paper. It also tells how important the accuracy is to analyze different machine learning algorithms that are used in the project.

Using business intelligence and open technologies, the authors provide a data analysis engine in their work [2]. For a more comprehensive picture of the problem, other data sources including metrological and traffic statistics are included into the air pollution study. Sensor networks in smart cities collect this information. The integrated data set is evaluated on a regular basis to produce informative dashboards using a selection of relevant performance metrics (KPIs). Using open data on air pollution in the urban area of a major station as a reference use case, the proposed technique is presented in a genuine smart city environment. So, this paper helps in knowing how to collect the information for the proposed model for the metropolitan cities. It also talks about how and where to collect the data.

Some of the world's most polluted cities are located in India and one of them is Delhi. Delhi's severe problems as a result of excessive air pollution are nothing new. To keep the public informed about the current and future state of air pollution, government organizations rely on air quality data. Thus, the Air Quality and Prediction is a very important and powerful methodology that helps us to forecast the air quality and analyze it. In the work first the collection of raw data of a particular city (Bangalore) is taken and then use the data preprocessing techniques for processing of data. In the next step splitting the data and testing the data are done using the sci-kit learn module. The Machine Learning Model is made such that the observed data can be generalized with trained data, so that when an unknown input is given the model should approximately give the output. The accuracy method is used to analyze the classifiers by using the training and testing data and find the best classifier after scoring.

LITERATURE SURVEY

In this paper we strived to achieve a better system that could give the best results of all available systems. During this process we have come across various findings of existing systems and have gone through various findings. We have studied findings of various researchers by studying various journals and research thesis and implemented various techniques using their studies and findings.

Research in the subject of Air quality has been ongoing over the past several years, with the goal of developing a flawless Air quality prediction system. Several papers have been published describing the need for Air quality prediction as well as the procedure for implementing various changes in the systems to provide more accuracy than in previous systems.

In this project a better system was achieved such that it could give the best results of all available systems. During this process various findings were come across of existing systems. The findings of various researchers were studied by studying various journals and research thesis and implemented various techniques using their studies and findings. Research in the subject of Air quality has been ongoing over the past several years, with the goal of developing a flawless Air quality prediction system. Several papers have been published describing the need for Air quality prediction as well as the procedure for implementing various changes in the systems to provide more accuracy than in previous systems. Many researchers have published various methods for overcoming various challenges through air quality prediction. Our research validates and reviews the publications that have laid the groundwork for the development of our thesis. The review explains what learning is done in this context by referring to various thesis and journals proposed by various research.

In the paper [3] the author researched in two distinct cities: Beijing and Italy. He used two publicly available datasets to forecast the AQI for Beijing and the NOx concentrations in Italian cities. He was right on both counts. The Beijing Municipal Environmental Center contributed the basic dataset of 1738 incidents from December 2013 to August 2018 including hourly average AQI, PM2.5, O3, SO2, PM10, and Beijing NO2. Italian towns were visited between March 2004 and February 2005 to gather the second round of 9358 cases. Benzene, non-methane volatiles, NOx, and NO2 concentrations are all included in this dataset. NOx prediction was our exclusive emphasis since it is one of the most significant air quality predictors. AQI and NOx concentrations were predicted using support vector regression (SVR) and random forest regression (RFR) approaches. The RFR and SVR are used for forecasting AQI and NOx concentration. The author focused on the NOx pollutant concentration which was one of the attribute in the proposed project. The author mentioned the NOx significance in this project that was very helpful for the proposed project.

The researcher in the paper [4] mentioned about the importance to look at the concentrations of various pollutants (NO, NO2, CO2, PM10, SO2) in the air at different times of day and throughout different zones of the city were estimated. The WEKA programme was used to calculate the velocity and humidity of the aforementioned air contaminants. The Karnataka Environmental Protection Agency was in charge of gathering the data. On weekdays, particularly during rush hour, air pollution levels rise, whereas on weekends and vacations they fall, as shown by the ZeroR algorithm and WEKA tool. K-means clustering is used to show the link between environmental parameters like temperature, wind velocity, and humidity and air pollutants like NO, NO2, PM10, CO, and SO2. The paper [4] was helpful for the proposed model as it was majorly discussing about various pollutants that were used as attributes and also how the K-means clustering was helpful in showing the link between the attributes.

The researcher of the paper [5] studied the air quality at Salem Swadeswari University in Tamilnadu from April 2011 to March 2011 and was the primary goal of this research. There are serious pollution problems associated with pollutants like gas, nitrogen oxides and suspended matter. The author mentioned about monitoring of data and how to monitor these type of data.

The author in this paper [6] does the analysis and prediction by using the python programming language. The inbuilt libraries and the environment that the author had worked on helps in learning and modelling the proposed work. The author helps in making the viewer's familiarize on python libraries and environment.

METHODOLOGY

The Air quality Analysis and prediction system designs contains 6 main steps as shown in Figure 2.

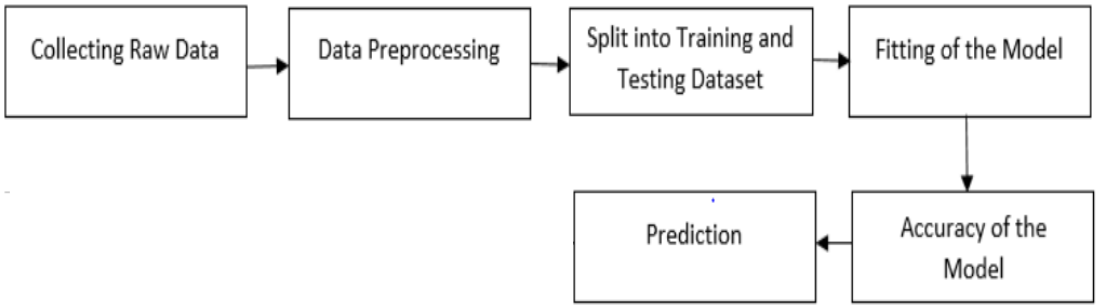

Figure 2: System Architecture

Collecting the Raw Data

Data collecting is the technique of aggregating and analysing data from various sources. Keeping track of previous events allows one to use data analysis to identify recurring patterns. Air Quality is a collection of statistics collected from the internet. The dataset included four independent variables (HC, NOX, SO, and HUMID) and one target variable (Pollution), all of which were measured as shown in Figure 3. Each attribute is represented with a value these are used as raw data.

```
HC,NOX,SO,HUMID,Pollution
21,15,59,59,1
8,10,39,57,1
6,6,33,54,1
18,8,24,56,2
43,38,206,55,2
30,32,72,54,2
21,32,62,56,3
6,4,4,56,3
18,12,37,61,3
12,7,20,59,1
21,15,59,59,1
8,10,39,57,1
6,6,33,54,1
18,8,24,56,2
43,38,206,55,2
30,32,72,54,2
21,32,62,56,3
6,4,4,56,3
18,12,37,61,3
12,7,20,59,1
21,15,59,59,1
8,10,39,57,1
6,6,33,54,1
18,8,24,56,2
43,38,206,55,2
30,32,72,54,2
21,32,62,56,3
6,4,4,56,3
18,12,37,61,3
```

Figure 3: Collecting raw data

Pre-processing

To transform raw data into a form that machine learning algorithms can utilize to gain insights or predict outcomes, the process of data pre-processing is termed. The data processing approach used in this project is the search for missing values. It's difficult to get all the data points for every record in a dataset. If there are any empty cells, fill the values for it. There were no missing values in the dataset utilized for the research.

Train and Test Split

The dataset is divided into a training dataset and a testing dataset using the train test split () technique in the sci-kit learn module. A training dataset (80% of the dataset) and a testing dataset (20% of the dataset).

Fitting the Model

The process of fine-tuning a model's parameters in order to improve its accuracy is called fitting. To construct a machine learning model, an algorithm is performed on data for which the target variable is already known. In order to assess the correctness of a model, it must be compared against real, observed values of the target variable.

When a machine learning model is able to generalize data that it was trained on, it is called model fitting. A good model fit is one that reliably predicts the result when given unknown inputs.

Accuracy of Model

Machine learning models may be used to predict the values of fresh input data, which is known as scoring. The degree to which a model has improved over time may be determined by calculating its accuracy score () using a training dataset.

Predicting the Model

"Prediction" is the term used to describe the outcome of an algorithm after it has been trained on a prior dataset and applied to fresh data. Using the test feature dataset to forecast the model with predict () method. It produced an array of expected values as the output.

Algorithms:

Improved KNN classifier is fed with Clustered Training Data D , Test samples T, k value, and Classes C. Classifier predicts the class label to each instance of test sample by following the below given steps.

Input : Clustered training data D, test samples T, k value and classes C.
Output: Test class prediction
Procedure: To each instance p in D
do
Compute $\text{lnd}(D(p1,p2)) = \log(\sum(\|pi\| - \|pj\|)^2)$;
done
Sort k-neighbors according to their distances.
sort(k, D(t,p))
Compute probability estimation for the sorted neighbors.
To each instance t(i) in k-neighbors(sort(k, D(t,p)))
do
$\text{DistProb}[] = \frac{1}{\sqrt{2.\pi}} \int e^{-D(t(i),p)^2} \, dD(t(i),p)/|N|; N = \text{Total attributes}$
done
To each test sample t in k-neighbors(sort(k, D(t,p)))
do
Compute class membership probabilities of each test sample t
assign class to t sample using classifier.
done

RESULTS

The raw data that we are collecting is from the source [9]. The Bangalore city data is taken and preprocessing the data is done. The data is splitted into training and testing data. Then accuracy for the model is calculated and the prediction for the dataset is derived.

First the credentials for User ID and password are written as admin to open the window. Then it will show three buttons each button has its own use as shown in Figure 4.

Figure 4: Admin Home window

For the Air Quality Analysis and AQI Prediction the Proposed classifier Improved KNN is found best out of the three classifiers like K-Nearest Neighbor, Multinomial Logistic Regression and Decision tree algorithm by calculating the accuracy score using the data. The result is shown as Figure 5:

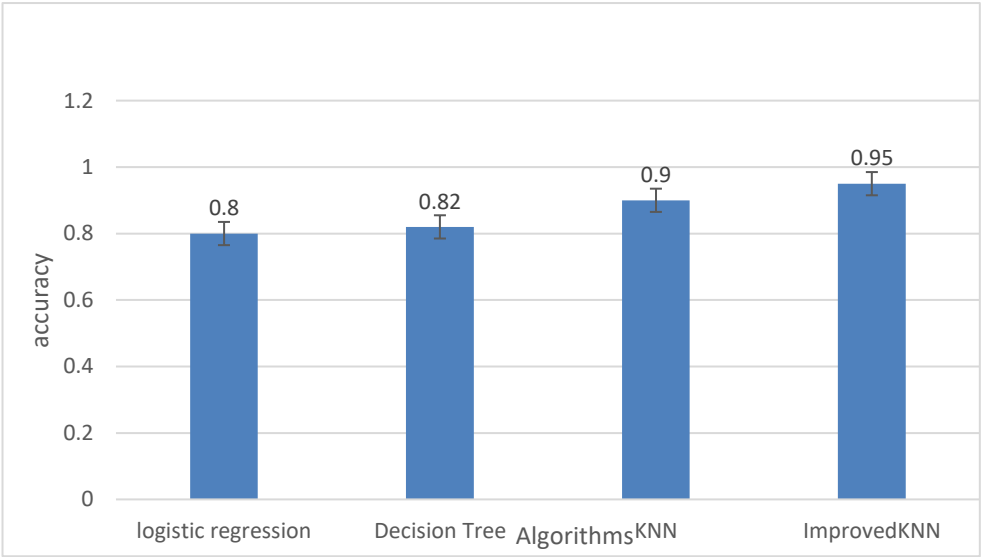

Figure 5: The result of the Air Quality Analysis

From the result from the three algorithms the K-Nearest Neighbor algorithm has the highest accuracy. Its accuracy is 0.95 that is 95% comparatively more with respect to other algorithms where decision tree has 80% and Multinomial logistic regression has 80%. So, this algorithm is used to predict the result whether it is good, moderate and unhealthy. The AQI ranges for these are as shown in Table 1:

AQI Category	Range

Good	0-50
Moderate	51-100
Unhealthy	>100

Table 1: AQI range

The test data is used and needed to be inputted in the .csv file as shown in Figure 6.

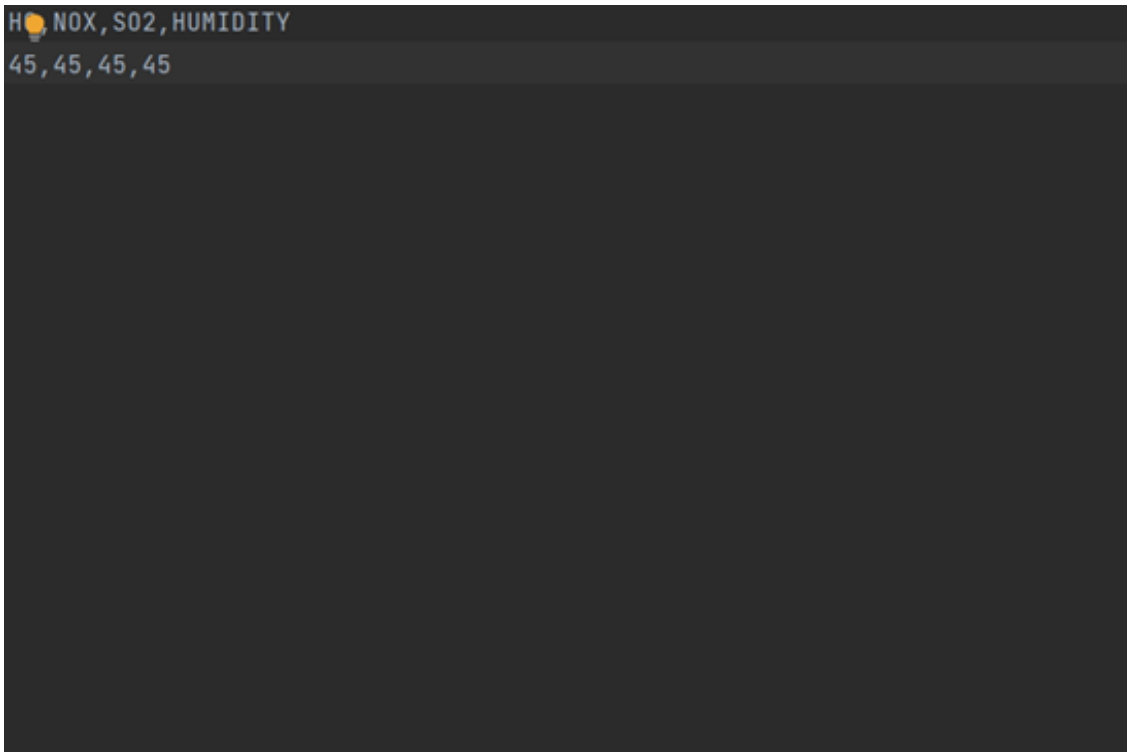

Figure 6: Input Data

Based on the input data the result is found out by comparing the similar data and classifying whether it is good, moderate, and unhealthy. For the data shown in Figure 6 it will show the result shown in Figure 7.

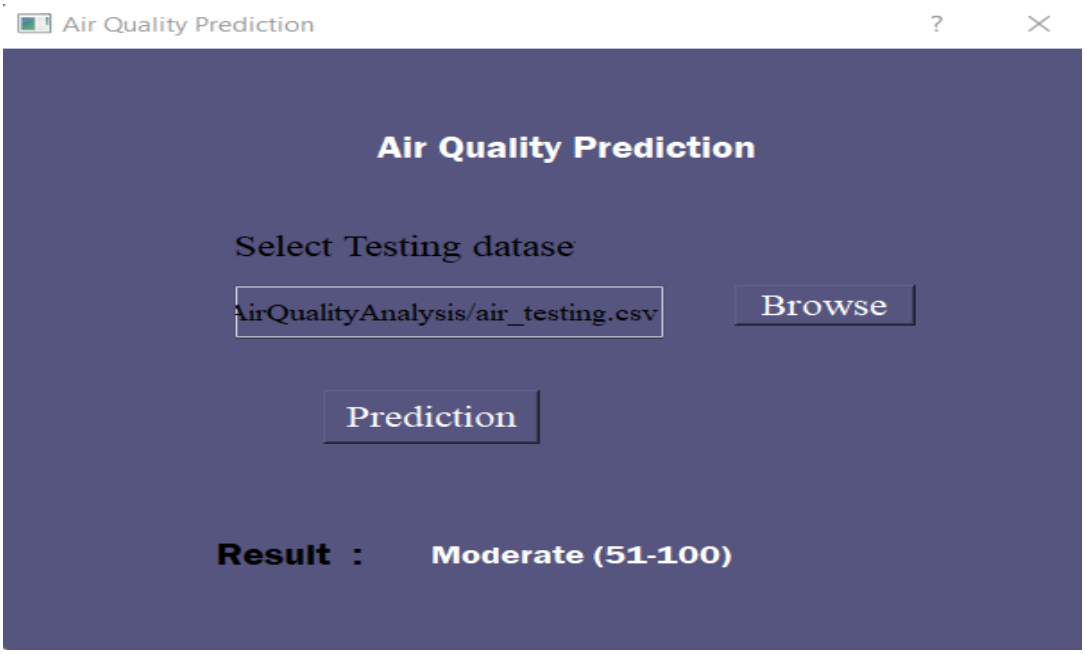

Figure 7: Prediction

The AQI value for the input is shown in the Figure 8.

Figure 8: AQI value

CONCLUSION AND FUTURE ENHANCEMENT

Predicting the quality of air is a complex task as there is a dynamic nature, high variability of pollutants and particulates. Air quality modelling and analysis is also becoming more relevant in metropolitan areas owing to the rise in human population and human-made activities and environmental change. The pollutants like HC, SO2, NOx, and Humidity are taken as the dataset from Bangalore in the government websites like Open Govt Data (OGD) and Central Pollution Control Board (CPCB) where daily raw data is provided in live, is preprocessed and transformed. In this project there are three algorithms that are compared like KNN, Decision Tree, and

Logistic Regression and finding the best classifier from this by using accuracy score. The K- Means algorithm is used for clustering the similar data in the project. The best classifier is used for predicting the result. Improved KNN classifier got the highest accuracy when comparing other classifiers in the project.

Research on Enhancing the Data to number of regions and experimenting on improving the cluster analysis will provide the best results in the future.

REFERENCES

[1] K. Kumar & B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities", International Journal of Environmental Science and Technology: IJEST 2022 May 15, 1-16.

[2] Elena Baralis, Tania Cerquitelli, Silvia Chiusano, "Analyzing air pollution on the urban environment" MIPRO 2016, May 30 - June 3, 2016, Opatija, Croatia.

[3] Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu, "Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms", Applied Sciences, ISSN 2076-3417; CODEN: ASPCC7, 2019, 9, 4069; doi:10.3390/app9194069.

[4] Mohamed Shakir, N. Rakesh, "Investigation on Air Pollutant Data Sets using Data Mining Tool", *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) IEEE Xplore*, Palladam, India, Part Number:CFP18OZV-ART; ISBN:978-1- 5386-1442-6.

[5] S. R. Gunasekaran, K. Kumaraswamy, P.P. Chandrasekaran, R. Elanchezhian, "MONITORING OF AMBIENT AIR QUALITY IN SALEM CITY, TAMIL NADU", *International Journal of Current Research(IJCR)*,Tamil Nadu, India, ISSN: 0975-833X, Vol. 4, Issue, 03, pp.275- 280, March, 2012.

[6] Yusef Omidi Khaniabadi, Gholamreza Goudarzi, Seyed Mohammad Daryanoosh, Alessandro Borgini, Andrea Tittarelli, Alessandra De Marco, "Exposure to PM10, NO2, and O3 and impacts on human health", *Environmental Science and Pollution Research,Springer* , volume 24, pp 2781–2789, Environ SciPollut Res, Berlin, Heidelberg, 2016.

[7] https://www.brainkart.com/article/Air-Pollution_38165/

[8] K. Nandini, G. Fathima. "Urban Air Quality Analysis and Prediction Using Machine Learning", *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, Bangalore, India, 2019.

[9] https://cpcb.nic.in/